

12th International Conference on Geographic Information Science

GIScience 2023, September 12–15, 2023, Leeds, UK

Edited by

Roger Beecham

Jed A. Long

Dianna Smith

Qunshan Zhao

Sarah Wise



Editors

Roger Beecham 

University of Leeds, GB
R.J.Beecham@leeds.ac.uk

Jed A. Long 

Western University, London, ON, Canada
jed.long@uwo.ca

Dianna Smith 

University of Southampton, GB
D.M.Smith@soton.ac.uk

Qunshan Zhao 

University of Glasgow, GB
qunshan.zhao@glasgow.ac.uk

Sarah Wise 

University College London, GB
s.wise@ucl.ac.uk

ACM Classification 2012

Information systems → Geographic information systems; Human-centered computing → Human computer interaction (HCI); Human-centered computing → Visualization; Theory of computation → Computational geometry; Computing methodologies → Machine learning; Information systems → Spatial-temporal systems

ISBN 978-3-95977-288-4

Published online and open access by

Schloss Dagstuhl – Leibniz-Zentrum für Informatik GmbH, Dagstuhl Publishing, Saarbrücken/Wadern, Germany. Online available at <https://www.dagstuhl.de/dagpub/978-3-95977-288-4>.

Publication date

September, 2023

Bibliographic information published by the Deutsche Nationalbibliothek

The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available in the Internet at <https://portal.dnb.de>.

License

This work is licensed under a Creative Commons Attribution 4.0 International license (CC-BY 4.0): <https://creativecommons.org/licenses/by/4.0/legalcode>.



In brief, this license authorizes each and everybody to share (to copy, distribute and transmit) the work under the following conditions, without impairing or restricting the authors' moral rights:

- Attribution: The work must be attributed to its authors.

The copyright is retained by the corresponding authors.

Digital Object Identifier: 10.4230/LIPIcs.GIScience.2023.0

ISBN 978-3-95977-288-4

ISSN 1868-8969

<https://www.dagstuhl.de/lipics>

LIPICs – Leibniz International Proceedings in Informatics

LIPICs is a series of high-quality conference proceedings across all fields in informatics. LIPICs volumes are published according to the principle of Open Access, i.e., they are available online and free of charge.

Editorial Board

- Luca Aceto (*Chair*, Reykjavik University, IS and Gran Sasso Science Institute, IT)
- Christel Baier (TU Dresden, DE)
- Roberto Di Cosmo (Inria and Université de Paris, FR)
- Faith Ellen (University of Toronto, CA)
- Javier Esparza (TU München, DE)
- Daniel Král' (Masaryk University, Brno, CZ)
- Meena Mahajan (Institute of Mathematical Sciences, Chennai, IN)
- Anca Muscholl (University of Bordeaux, FR)
- Chih-Hao Luke Ong (University of Oxford, GB)
- Phillip Rogaway (University of California, Davis, US)
- Eva Rotenberg (Technical University of Denmark, Lyngby, DK)
- Raimund Seidel (Universität des Saarlandes, Saarbrücken, DE and Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Wadern, DE)
- Pierre Senellart (ENS, Université PSL, Paris, FR)

ISSN 1868-8969

<https://www.dagstuhl.de/lipics>

■ Contents

Preface	
<i>Roger Beecham, Jed A. Long, Dianna Smith, Qunshan Zhao, and Sarah Wise</i>	0:xiii
Authors	
.....	0:xv–0:xxiv

Regular Papers

Do You Need Instructions Again? Predicting Wayfinding Instruction Demand	
<i>Negar Alinaghi, Tiffany C. K. Kwok, Peter Kiefer, and Ioannis Giannopoulos</i>	1:1–1:16
Transitions in Dynamic Point Labeling	
<i>Thomas Depian, Guangping Li, Martin Nöllenburg, and Jules Wolms</i>	2:1–2:19
Reducing False Discoveries in Statistically-Significant Regional-Colocation Mining: A Summary of Results	
<i>Subhankar Ghosh, Jayant Gupta, Arun Sharma, Shuai An, and Shashi Shekhar</i> ..	3:1–3:18
Genetic Programming for Computationally Efficient Land Use Allocation Optimization	
<i>Moritz J. Hildemann, Alan T. Murray, and Judith A. Versteegen</i>	4:1–4:15
Visualizing Geophylogenies – Internal and External Labeling with Phylogenetic Tree Constraints	
<i>Jonathan Klawitter, Felix Klesen, Joris Y. Scholl, Thomas C. van Dijk, and Alexander Zaft</i>	5:1–5:16
Map Reproducibility in Geoscientific Publications: An Exploratory Study	
<i>Eftychia Koukouraki and Christian Kray</i>	6:1–6:16
Semi-Supervised Learning from Street-View Images and OpenStreetMap for Automatic Building Height Estimation	
<i>Hao Li, Zhendong Yuan, Gabriel Dax, Gefei Kong, Hongchao Fan, Alexander Zipf, and Martin Werner</i>	7:1–7:15
Towards a Multidimensional Interaction Framework for Promoting Public Engagement in Citizen Science Projects	
<i>Maryam Lotfian, Jens Ingensand, and Christophe Claramunt</i>	8:1–8:16
Platial k-Anonymity: Improving Location Anonymity Through Temporal Popularity Signatures	
<i>Grant McKenzie and Hongyu Zhang</i>	9:1–9:15
Data-Spatial Layouts for Grid Maps	
<i>Nathan van Beusekom, Wouter Meulemans, Bettina Speckmann, and Jo Wood</i> ...	10:1–10:17
Benchmarking Regression Models Under Spatial Heterogeneity	
<i>Nina Wiedemann, Henry Martin, and René Westerholt</i>	11:1–11:15

12th International Conference on Geographic Information Science (GIScience 2023).

Editors: Roger Beecham, Jed A. Long, Dianna Smith, Qunshan Zhao, and Sarah Wise

Leibniz International Proceedings in Informatics



LIPIC Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Short Papers

Confidential, Decentralized Location-Based Data Services <i>Benjamin Adams</i>	12:1–12:6
Towards an Inclusive Urban Environment: A Participatory Approach for Collecting Spatial Accessibility Data in Zurich <i>Hoda Allahbakhshi</i>	13:1–13:6
Development of a Semantic Segmentation Approach to Old-Map Comparison <i>Yves Annanias, Daniel Wiegrefe, Andreas Niekler, Marta Kuźma, and Francis Harvey</i>	14:1–14:6
Why Is Greenwich so Common? Quantifying the Uniqueness of Multivariate Observations <i>Andrea Ballatore and Stefano Cavazzi</i>	15:1–15:6
When Everything Is “Nearby”: How Airbnb Listings in New York City Exaggerate Proximity <i>Mikael Brunila, Priyanka Verma, and Grant McKenzie</i>	16:1–16:8
Smarter Than Your Average Model - Bayesian Model Averaging as a Spatial Analysis Tool <i>Chris Brunsdon, Paul Harris, and Alexis Comber</i>	17:1–17:6
Anonymous Routing Using Minimum Capacity Clustering <i>Maike Buchin and Lukas Plätz</i>	18:1–18:6
Achieving Least Relocation of Existing Facilities in Spatial Optimisation: A Bi-Objective Model <i>Huanfa Chen and Rongbo Xu</i>	19:1–19:5
Exploring Energy Deprivation Across Small Areas in England and Wales <i>Meixu Chen, Alex Singleton, and Caitlin Robinson</i>	20:1–20:6
Using the Dynamic Microsimulation MINOS to Evidence the Effect of Energy Crisis Income Support Policy <i>Robert Clay, Luke Archer, Alison Heppenstall, and Nik Lomax</i>	21:1–21:6
Multiscale Spatially and Temporally Varying Coefficient Modelling Using a Geographic and Temporal Gaussian Process GAM (GTGP-GAM) <i>Alexis Comber, Paul Harris, and Chris Brunsdon</i>	22:1–22:6
Does Generalisation Matter in Pan-Scalar Maps? <i>Azelle Courtial and Guillaume Touya</i>	23:1–23:6
Understanding People’s Perceptions of Their Liveable Neighbourhoods: A Case Study of East Bristol <i>Elisa Covato and Shelan Jeawak</i>	24:1–24:6
Building Alternative Indices of Socioeconomic Status for Population Modeling in Data-Sparse Contexts <i>Angela R. Cunningham, Joseph V. Tuccillo, and Tyler J. Frazier</i>	25:1–25:7
Uncertainty in Causal Neighborhood Effects: A Multi-Agent Simulation Approach <i>Cécile de Bézenac</i>	26:1–26:6

Uncovering Spatiotemporal Patterns of Travel Flows Under Extreme Weather Events by Tensor Decomposition <i>Zhicheng Deng, Zhaoya Gong, and Pengjun Zhao</i>	27:1–27:6
GeoQAMap - Geographic Question Answering with Maps Leveraging LLM and Open Knowledge Base <i>Yu Feng, Linfang Ding, and Guohui Xiao</i>	28:1–28:7
Understanding the Complex Behaviours of Electric Vehicle Drivers with Agent-Based Models in Glasgow <i>Zixin Feng, Qunshan Zhao, and Alison Heppenstall</i>	29:1–29:6
Progress in Constructing an Open Map Generalization Data Set for Deep Learning <i>Cheng Fu, Zhiyong Zhou, Jan Winkler, Nicolas Beglinger, and Robert Weibel</i>	30:1–30:6
Project-Based Urban Dynamics: A Novel Method for Assessing Urban Sprawl <i>Nir Fulman, Yulia Grinblat, and Itzhak Benenson</i>	31:1–31:6
From Reproducible to Explainable GIScience <i>Mark Gahegan</i>	32:1–32:6
Uncertainty Quantification in the Road-Level Traffic Risk Prediction by Spatial-Temporal Zero-Inflated Negative Binomial Graph Neural Network(STZINB-GNN) <i>Xiaowei Gao, James Haworth, Dingyi Zhuang, Huanfa Chen, and Xinke Jiang</i> ...	33:1–33:6
Simulating and Validating the Traffic of Blackwall Tunnel Using TFL Jam Cam Data and Simulation of Urban Mobility (SUMO) <i>Chukun Gao</i>	34:1–34:8
Building-Level Comparison of Microsoft and Google Open Building Footprints Datasets <i>Jack Joseph Gonzales</i>	35:1–35:6
Characterizing Urban Expansion Processes Using Dynamic Spatial Models – a European Application <i>Alex Hagen-Zanker, Jingyan Yu, Naratip Santitissadeekorn, and Susan Hughes</i> ...	36:1–36:6
Understanding the Spatial Complexity in Landscape Narratives Through Qualitative Representation of Space <i>Erum Haris, Anthony G. Cohn, and John G. Stell</i>	37:1–37:6
Exascale Agent-Based Modelling for Policy Evaluation in Real-Time (ExAMPLER) <i>Alison Heppenstall, J. Gary Polhill, Mike Batty, Matt Hare, Doug Salt, and Richard Milton</i>	38:1–38:5
A Hierarchical and Geographically Weighted Regression Model and Its Backfitting Maximum Likelihood Estimator <i>Yigong Hu, Richard Harris, Richard Timmerman, and Binbin Lu</i>	39:1–39:6
Introducing a General Framework for Locally Weighted Spatial Modelling Based on Density Regression <i>Yigong Hu, Binbin Lu, Richard Harris, and Richard Timmerman</i>	40:1–40:7

Understanding Place Identity with Generative AI <i>Kee Moon Jang, Junda Chen, Yuhao Kang, Junghwan Kim, Jinhjung Lee, and Fábio Duarte</i>	41:1–41:6
An Integrated Uncertainty and Sensitivity Analysis for Spatial Multicriteria Models <i>Piotr Jankowski, Arika Ligmann-Zielińska, Zbigniew Zwoliński, and Alicja Najwer</i>	42:1–42:6
Evaluating the Effectiveness of Large Language Models in Representing Textual Descriptions of Geometry and Spatial Relations <i>Yuhan Ji and Song Gao</i>	43:1–43:6
Framework for Motorcycle Risk Assessment Using Onboard Panoramic Camera <i>Natchapon Jongwirayanurak, Zichao Zeng, Meihui Wang, James Haworth, Garavig Tanaksaranond, and Jan Boehm</i>	44:1–44:7
National-Scale Spatiotemporal Variation in Driver Navigation Behaviour and Route Choice <i>Elliot Karikari, Manon Prédhumeau, Peter Baudains, and Ed Manley</i>	45:1–45:6
Status Poles and Status Zoning to Model Residential Land Prices: Status-Quality Trade off Theory <i>Thuy Phuong Le, Alexis Comber, Binh Quoc Tran, Phe Huu Hoang, Huy Quang Man, Linh Xuan Nguyen, Tuan Le Pham, and Tu Ngoc Bui</i>	46:1–46:6
Investigating MAUP Effects on Census Data Using Approximately Equal-Population Aggregations <i>Yue Lin and Ningchuan Xiao</i>	47:1–47:6
Agent-Based Modelling and Disease: Demonstrating the Role of Human Remains in Epidemic Outbreaks <i>Huixin Liu and Sarah Wise</i>	48:1–48:7
How Does Travel Environment Affect Mood? A Study Using Geographic Ecological Momentary Assessment in the UK <i>Milad Malekzadeh, Darja Reuschke, and Jed A. Long</i>	49:1–49:6
Calibration in a Data Sparse Environment: How Many Cases Did We Miss? <i>Robert Manning Smith, Sarah Wise, and Sophie Ayling</i>	50:1–50:7
Geographic Analysis of Trade-Offs Between Amenity and Supply Effects in New Office Buildings <i>Kazushi Matsuo, Morito Tsutsumi, and Toyokazu Imazeki</i>	51:1–51:6
Impacts of Catchments Derived from Fine-Grained Mobility Data on Spatial Accessibility <i>Alexander Michels, Jinwoo Park, Bo Li, Jeon-Young Kang, and Shaowen Wang</i> ..	52:1–52:6
Exploring the Potential of Machine and Deep Learning Models for OpenStreetMap Data Quality Assessment and Improvement <i>Salim Miloudi and Bouhadjar Meguenni</i>	53:1–53:6
On the Cartographic Communication of Places <i>Franz-Benjamin Mocnik</i>	54:1–54:6

Resiliency: A Consensus Data Binning Method <i>Arpit Narechania, Alex Endert, and Clio Andris</i>	55:1–55:7
Counter-Intuitive Effect of Null Hypothesis on Moran’s <i>I</i> Tests Under Heterogenous Populations <i>Hayato Nishi and Ikuho Yamada</i>	56:1–56:6
A Data Fusion Framework for Exploring Mobility Around Disruptive Events <i>Evgeny Noi and Somayeh Dodge</i>	57:1–57:6
Finding Feasible Routes with Reinforcement Learning Using Macro-Level Traffic Measurements <i>Mustafa Can Ozkan and Tao Cheng</i>	58:1–58:6
Moran Eigenvectors-Based Spatial Heterogeneity Analysis for Compositional Data <i>Zhan Peng and Ryo Inoue</i>	59:1–59:6
Toward Causally Aware GIS: Events as Cornerstones <i>Nina Polous</i>	60:1–60:8
Mobility Vitality: Assessing Neighborhood Similarity Through Transportation Patterns In New York City <i>Dan Qiang and Grant McKenzie</i>	61:1–61:6
An Evaluation of the Impact of Ignition Location Uncertainty on Forest Fire Ignition Prediction Using Bayesian Logistic Regression <i>David Röbl, Rizwan Bulbul, Johannes Scholz, Mortimer M. Müller, and Harald Vacik</i> . 62:1–62:7	
Calculating Shadows with U-Nets for Urban Environments <i>Dominik Rothschedl, Franz Welscher, Franziska Hübl, Ivan Majic, Daniele Giannandrea, Matthias Wastian, Johannes Scholz, and Niki Popper</i>	63:1–63:6
Beware the Rise of Models When They Are Wrong: A Look at Heat Vulnerability Modeling Through the Lens of Sensitivity <i>Seda Şalap-Ayça and Erica Akemi Goto</i>	64:1–64:6
From Change Detection to Change Analytics: Decomposing Multi-Temporal Pixel Evolution Vectors <i>Victoria Scherelis, Patrick Laube, and Michael Doering</i>	65:1–65:6
How to Count Travelers Without Tracking Them Between Locations <i>Nadia Shafaeipour, Maarten van Steen, and Frank O. Ostermann</i>	66:1–66:6
A Personalised Pedestrian Navigation System <i>Urmi Shah and Jia Wang</i>	67:1–67:6
Estimating the Impact of a Flood Event on Property Value and Its Diminished Effect over Time <i>Nazia Ferdause Sodial, Oleksandr Galkin, and Aidan Slingsby</i>	68:1–68:6
Development and Operationalisation of Local Sustainability Indicators - A Global South Perspective on Data Challenges and Opportunities for GIScience <i>Stefan Steiniger, Carolina Rojas, Ricardo Truffello, and Jonathan Barton</i>	69:1–69:6

Assessing Epidemic Spreading Potential with Encounter Network <i>Behnam Tahmasbi, Farnoosh Roozkhosh, and X. Angela Yao</i>	70:1–70:6
Inferring the History of Spatial Diffusion Processes <i>Takuya Takahashi, Geneviève Hannes, Nico Neureiter, and Peter Ranacher</i>	71:1–71:6
Modelling Affordances as Emergent Phenomena <i>Sabine Timpf and Franziska Klügl</i>	72:1–72:6
The FogDetector: A User Survey to Measure Disorientation in Pan-Scalar Maps <i>Guillaume Touya and Justin Berli</i>	73:1–73:6
An Interpretable Index of Social Vulnerability to Environmental Hazards <i>Joseph V. Tuccillo</i>	74:1–74:6
Power of GIS Mapping: ATLAS Flood Maps 2022 <i>Munazza Usmani, Hafiz Muhammad Tayyab Bhatti, Francesca Bovolo, and Maurizio Napolitano</i>	75:1–75:6
A Data-Driven Decision-Making Framework for Spatial Agent-Based Models of Infectious Disease Spread <i>Emma Von Hoene, Amira Roess, and Taylor Anderson</i>	76:1–76:7
How to Improve Joint Suitability Mapping for Search Space Reduction? <i>Haoyu Wang and Jennifer A. Miller</i>	77:1–77:6
Navigation in Complex Space: An Bayesian Nash Equilibrium-Informed Agent-Based Model <i>Yiyu Wang, Jiaqi Ge, and Alexis Comber</i>	78:1–78:6
Application of GIS in Public Health Practice: A Consortium’s Approach to Tackling Travel Delays in Obstetric Emergencies in Urban Areas <i>Jia Wang, Itohan Osayande, Peter M. Macharia, Prestige Tatenda Makanga, Kerry L. M. Wong, Tope Olubodun, Uchenna Gwacham-Anisiobi, Olakunmi Ogunyemi, Abimbola Olaniran, Ibukun-Oluwa O. Abejirinde, Lenka Beňová, Bosede B. Afolabi, and Aduragbemi Banke-Thomas</i>	79:1–79:6
The Ups and Downs of London High Streets Throughout COVID-19 Pandemic: Insights from Footfall-Based Clustering Analysis <i>Xinglei Wang, Xianghui Zhang, and Tao Cheng</i>	80:1–80:6
Agent-Based Modeling of Consumer Choice by Utilizing Crowdsourced Data and Deep Learning <i>Boyu Wang and Andrew Crooks</i>	81:1–81:6
Harnessing the Sunlight on Facades – an Approach for Determining Vertical Photovoltaic Potential <i>Franz Welscher, Ivan Majic, Franziska Hübl, Rizwan Bulbul, and Johannes Scholz</i>	82:1–82:7
Betweenness Centrality in Spatial Networks: A Spatially Normalised Approach <i>Christian Werner and Martin Loidl</i>	83:1–83:6
Predicting visit frequencies to new places <i>Nina Wiedemann, Ye Hong, and Martin Raubal</i>	84:1–84:6

Waffle Homes: Utilizing Aerial Imagery of Unfinished Buildings to Determine Average Room Size <i>Carson Woody and Tyler Frazier</i>	85:1–85:6
A Comparison of Global and Local Statistical and Machine Learning Techniques in Estimating Flash Flood Susceptibility <i>Jing Yao, Ziqi Li, Xiaoxiang Zhang, Changjun Liu, and Liliang Ren</i>	86:1–86:6
Understand the Geography of Financial Precarity in England and Wales <i>Zi Ye and Alex Singleton</i>	87:1–87:6
Understanding Active Travel Networks Using GPS Data from an Outdoor Mapping App <i>Marcus A. Young</i>	88:1–88:6
Geography and the Brain’s Spatial System <i>May Yuan and Kristen Kennedy</i>	89:1–89:7
Visual Methods for Representing Flow Space with Vector Fields <i>Han Zhang, Zhaoya Gong, and Jean-Claude Thill</i>	90:1–90:6
Causal Effects Under Spatial Confounding and Interference <i>Jing Zhang</i>	91:1–91:6
Unlocking the Power of Mobile Phone Application Data to Accelerate Transport Decarbonisation <i>Xianghui Zhang and Tao Cheng</i>	92:1–92:6
The Ethics of AI-Generated Maps: DALL · E 2 and AI’s Implications for Cartography <i>Qianheng Zhang, Yuhao Kang, and Robert Roth</i>	93:1–93:6
Digital Injustice: A Case Study of Land Use Classification Using Multisource Data in Nairobi, Kenya <i>Wenlan Zhang, Chen Zhong, and Faith Taylor</i>	94:1–94:6
Exploring Map App Usage Behaviour Through Touchscreen Interactions <i>Donatella Zingaro, Mona Bartling, and Tumasch Reichenbacher</i>	95:1–95:6

■ Preface

This volume contains paper proceedings of the 12th International Conference on Geographic Information Science (GIScience 2023), held at University of Leeds in collaboration with University of Glasgow, 12–15 September 2023.

The conference attracted a large number of high quality submissions. Each submitted paper received three to four reviews plus a summary review, and in total we accepted 11 long papers and 84 short papers. In addition to these papers, 60 poster lightning talks were presented at the conference.

The accepted papers represent a wide range of topics at the forefront of GIScience research including work on spatial networks, movement analysis, agent-based modelling, spatial modelling and statistics, new forms of data and GeoAI, uncertainty representation, reproducibility, as well as papers relating to the conference theme – Disrupting Society. Separate to these contributions to the main conference was a suite of 14 pre-conference workshops hosting talks, tutorials, and panel discussions on a wide range of topics in GIScience.

The entire GIScience 2023 team would like to express their gratitude to the authors, reviewers, workshop and tutorial organisers, and everyone else involved in the conference.



■ List of Authors

Ibukun-Oluwa O. Abejirinde (79)

Dalla Lana School of Public Health, University of Toronto, Canada; Women's College Hospital Institute for, Health System Solutions and Virtual Care, Toronto, Canada

Benjamin Adams  (12)

Computer Science and Software Engineering, University of Canterbury, Christchurch, New Zealand

Bosede B. Afolabi (79)

Department of Obstetrics and Gynaecology, College of Medicine, University of Lagos, Nigeria; Maternal and Reproductive Health Research , Collective, Lagos, Nigeria

Negar Alinaghi  (1)

Geoinformation, TU Wien, Austria

Hoda Allahbakhshi (13)

Digital Society Initiative, University of Zürich, Switzerland; Department of Geography, University of Zürich, Switzerland

Shuai An (3)

Department of Economics, University of Minnesota, Minneapolis, MN, USA

Taylor Anderson  (76)

Department of Geography and Geoinformation Science, George Mason University, Fairfax, VA, USA

Clio Andris  (55)

Georgia Institute of Technology, Atlanta, GA, USA

Yves Annanias  (14)

Image and Signal Processing Group, Leipzig University, Germany

Luke Archer  (21)

University of Leeds, UK

Sophie Ayling  (50)


The Bartlett Centre for Advanced Spatial Analysis, University College London, UK

Andrea Ballatore  (15)

Department of Digital Humanities, King's College London, UK

Aduragbemi Banke-Thomas (79)

Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, UK; Maternal and Reproductive Health Research, Collective, Lagos, Nigeria; School of Human Sciences, University of Greenwich, London, UK

Mona Bartling  (95)


Department of Geography, University of Zurich, Switzerland

Jonathan Barton  (69)

Pontificia Universidad Católica de Chile, Santiago, Chile; CEDEUS, Santiago, Chile

Mike Batty  (38)

Bartlett Centre for Advanced Spatial Analysis, University College London, UK

Peter Baudains  (45)

ESRC Consumer Data Research Centre, University of Leeds, UK

Nicolas Beglinger (30)

swisstopo, Swiss Federal Office of Topography, Wabern, Switzerland

Itzhak Benenson (31)

Department of Geography and Human Environment, Porter School of Environmental Studies, Tel Aviv University, Israel

Justin Berli (73)

LASTIG, Univ Gustave Eiffel, IGN-ENSG, F-77420 Champs-sur-Marne, France

Lenka Beňová (79)

Department of Public Health, Institute of Tropical Medicine, Antwerp, Belgium

Hafiz Muhammad Tayyab Bhatti (75)

University of Punjab, Lahore, Pakistan

Jan Boehm  (44)

Department of Civil, Environmental and Geomatic Engineering, University College London, UK

Francesca Bovolo (75)

Fondazione Bruno Kessler, Trento, Italy

Mikael Brunila  (16)

Platial Analysis Lab, Department of Geography, McGill University, Montréal, Canada; Urban Politics & Governance Lab, School of Urban Planning, McGill University, Montréal, Canada

12th International Conference on Geographic Information Science (GIScience 2023).

Editors: Roger Beecham, Jed A. Long, Dianna Smith, Qunshan Zhao, and Sarah Wise




Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

- Chris Brunsdon  (17, 22)
National Centre for Geocomputation, Maynooth University, Ireland
- Maike Buchin (18)
Ruhr University Bochum, Germany
- Tu Ngoc Bui (46)
VNU University of Science, Hanoi, Vietnam
- Rizwan Bulbul  (62, 82)
Institute of Geodesy, Graz University of Technology, Graz, Austria
- Stefano Cavazzi  (15)
Ordnance Survey, Southampton, UK
- Huanfa Chen  (19, 33)
Centre for Advanced Spatial Analysis, University College London, UK
- Junda Chen (41)
DataChat, Madison, WI, USA
- Meixu Chen  (20)
Department of Geography and Planning, University of Liverpool, UK
- Tao Cheng  (58, 80, 92)
SpaceTimeLab, University College London, UK
- Christophe Claramunt  (8)
Naval Academy Research Institute, Brest Naval, Lanveoc-Poulmic, BP 600, 29240 Brest Naval, France
- Robert Clay  (21)
University of Leeds, UK
- Anthony G. Cohn  (37)
School of Computing, University of Leeds, UK; The Alan Turing Institute, London, UK
- Alexis Comber  (17, 22, 46, 78)
School of Geography, University of Leeds, UK
- Azelle Courtial  (23)
LASTIG, Univ Gustave Eiffel, IGN-ENSG, F-77420 Champs-sur-Marne, France
- Elisa Covato (24)
School of Computer Science and Creative Technologies, University of the West of England, Bristol, UK
- Andrew Crooks  (81)
Department of Geography, University at Buffalo, NY, USA
- Angela R. Cunningham  (25)
Oak Ridge National Laboratory, TN, USA
- Gabriel Dax (7)
Technical University of Munich, Germany
- Cécile de Bézenac (26)
University of Leeds, UK; The Alan Turing Institute, London, UK
- Zhicheng Deng  (27)
School of Urban Planning and Design, Peking University Shenzhen Graduate School, Shenzhen, Guangdong, China; Key Laboratory of Earth Surface System and Human-Earth Relations of Ministry of Natural Resources of China, Peking University Shenzhen Graduate School, Shenzhen, Guangdong, China
- Thomas Depian (2)
Algorithms and Complexity Group, TU Wien, Austria
- Linfang Ding  (28)
Norwegian University of Science and Technology, Trondheim, Norway
- Somayeh Dodge  (57)
Department of Geography, University of California Santa Barbara, CA, USA
- Michael Doering  (65)
Zurich University of Applied Sciences, Wädenswil, Switzerland
- Fábio Duarte  (41)
MIT Senseable City Lab, Cambridge, MA, USA
- Alex Endert  (55)
Georgia Institute of Technology, Atlanta, GA, USA
- Hongchao Fan (7)
Norwegian University of Science and Technology, Trondheim, Norway
- Yu Feng  (28)
Chair of Cartography and Visual Analytics, Technical University of Munich, Germany
- Zixin Feng (29)
Urban Big Data Centre, School of Social and Political Sciences, University of Glasgow, UK
- Tyler Frazier (85)
Human Geography Group, Oak Ridge National Laboratory, TN, USA
- Tyler J. Frazier  (25)
Oak Ridge National Laboratory, TN, USA
- Cheng Fu  (30)
Department of Geography, University of Zürich, Switzerland


- Nir Fulman (31)
Department of Geography and Human Environment, Porter School of Environmental Studies, Tel Aviv University, Israel; GIScience Research Group, Heidelberg University, Germany
- Mark Gahegan (32)
School of Computer Science / Centre for eResearch, University of Auckland, New Zealand
- Oleksandr Galkin (68)
City, University of London, UK
- Chukun Gao (34)
Centre for Advanced Spatial Analysis, University College London, UK
- Song Gao  (43)
GeoDS Lab, Department of Geography, University of Wisconsin-Madison, WI, USA
- Xiaowei Gao  (33)
SpaceTimeLab, University College London (UCL), UK
- Jiaqi Ge  (78)
School of Geography, University of Leeds, UK
- Subhankar Ghosh (3)
Department of Computer Science & Engineering, University of Minnesota, Minneapolis, MN, USA
- Daniele Giannandrea (63)
dwh GmbH, Vienna, Austria; Institute of Information Systems Engineering, TU Vienna, Austria
- Ioannis Giannopoulos  (1)
Geoinformation, TU Wien, Austria, Institute of Advanced Research in Artificial Intelligence (IARAI), Austria
- Zhaoya Gong  (27, 90)
School of Urban Planning and Design, Peking University Shenzhen Graduate School, Shenzhen, Guangdong, China; Key Laboratory of Earth Surface System and Human-Earth Relations of Ministry of Natural Resources of China, Peking University Shenzhen Graduate School, Shenzhen, Guangdong, China
- Jack Joseph Gonzales  (35)
Geospatial Science and Human Security Division, Oak Ridge National Laboratory, TN, USA
- Erica Akemi Goto  (64)
Arizona Institute for Resilience, University of Arizona, Tucson, AZ, USA
- Yulia Grinblat (31)
Heidelberg Institute for Geoinformation Technology (HeiGIT) gGmbH at Heidelberg University, Germany
- Jayant Gupta (3)
Department of Computer Science & Engineering, University of Minnesota, Minneapolis, MN, USA
- Uchenna Gwacham-Anisiobi (79)
Nuffield Department of Population Health, University of Oxford, UK
- Alex Hagen-Zanker  (36)
School of Sustainability, Civil and Environmental Engineering, University of Surrey, UK
- Geneviève Hannes  (71)
Department of Geography, University of Zurich, Switzerland
- Matt Hare  (38)
The James Hutton Institute, Aberdeen, UK
- Erum Haris  (37)
School of Computing, University of Leeds, UK
- Paul Harris  (17, 22)
Rothamsted Research, Harpenden, UK
- Richard Harris  (39, 40)
School of Geographical Sciences, University of Bristol, UK
- Francis Harvey  (14)
Leibniz Institute for Regional Geography, Leipzig, Germany; Faculty of History, University of Warsaw, Poland
- James Haworth  (33, 44)
SpaceTimeLab, University College London (UCL), UK
- Alison Heppenstall  (21, 29, 38)
School of Political and Social Sciences, MRC/CSO Social and Public Health Sciences Unit, University of Glasgow, UK
- Moritz J. Hildemann  (4)
Institute for Geoinformatics, University of Münster, Germany
- Phe Huu Hoang (46)
R & D Consultants, Hanoi, Vietnam


- Ye Hong  (84)
Institute of Cartography and Geoinformation,
ETH Zürich, Switzerland
- Yigong Hu  (39, 40)
School of Geographical Sciences, University of
Bristol, UK
- Susan Hughes  (36)
School of Sustainability, Civil and
Environmental Engineering, University of Surrey,
UK
- Franziska Hübl  (63, 82)
Institute of Geodesy, Graz University of
Technology, Austria
- Toyokazu Imazeki  (51)
Commercial Property Research Institute, Inc.,
Tokyo, Japan
- Jens Ingensand  (8)
Institute INSIT, School of Business and
Engineering Vaud, University of Applied
Sciences and Arts Western Switzerland,
Yverdon-les-Bains, Switzerland
- Ryo Inoue  (59)
Graduate School of Information Sciences,
Tohoku University, Sendai, Japan
- Kee Moon Jang  (41)
MIT Senseable City Lab, Cambridge, MA, USA
- Piotr Jankowski  (42)
San Diego State University, CA, USA; Adam
Mickiewicz University, Poznan, Poland
- Shelan Jeawak (24)
School of Computer Science and Creative
Technologies, University of the West of England,
Bristol, UK
- Yuhan Ji  (43)
GeoDS Lab, Department of Geography,
University of Wisconsin-Madison, WI, USA
- Xinke Jiang  (33)
School of Computer Science, Peking University
(PKU), Beijing, China
- Natchapon Jongwiriyanurak  (44)
Department of Civil, Environmental and
Geomatic Engineering, University College
London, UK
- Jeon-Young Kang  (52)
Department of Geography, Kyung Hee
University, Dongdaemun-gu, Seoul, South Korea
- Yuhao Kang  (41, 93)
MIT Senseable City Lab, Cambridge, MA, USA
- Elliot Karikari  (45)
Leeds Institute for Data Analytics, University of
Leeds, UK
- Kristen Kennedy  (89)
Cognition and Neuroscience, The University of
Texas at Dallas, TX, USA
- Peter Kiefer  (1)
Institute of Cartography and Geoinformation,
ETH Zürich, Switzerland
- Junghwan Kim  (41)
Department of Geography, Virginia Tech,
Blacksburg, VA, USA
- Jonathan Klawitter  (5)
University of Auckland, New Zealand
- Felix Klesen  (5)
Universität Würzburg, Germany
- Franziska Klügl  (72)
AASS/NT, Örebro University, Sweden
- Gefei Kong (7)
Norwegian University of Science and Technology,
Trondheim, Norway
- Eftychia Koukouraki  (6)
Institute for Geoinformatics, University of
Münster, Germany
- Christian Kray  (6)
Institute for Geoinformatics, University of
Münster, Germany
- Marta Kuźma  (14)
Faculty of History, University of Warsaw,
Poland
- Tiffany C. K. Kwok  (1)
Institute of Cartography and Geoinformation,
ETH Zürich, Switzerland
- Patrick Laube  (65)
Zurich University of Applied Sciences,
Wädenswil, Switzerland; University of Zurich,
Switzerland
- Tuan Le Pham (46)
VNU University of Science, Hanoi, Vietnam
- Thuy Phuong Le  (46)
VNU University of Science, Hanoi, Vietnam
- Jinhyung Lee  (41)
Department of Geography and Environment,
Western University, London, Canada


Bo Li (52)
Department of Statistics, University of Illinois
Urbana-Champaign, IL, USA

Guangping Li  (2)
Algorithm Engineering Group, TU Dortmund,
Germany

Hao Li (7)
Technical University of Munich, Germany

Ziqi Li  (86)
Department of Geography, Florida State
University, Tallahassee, FL, USA


Arika Ligmann-Zielińska  (42)
Michigan State University, East Lansing, MI,
USA; Adam Mickiewicz University, Poznan,
Poland


Yue Lin  (47)
Department of Geography, The Ohio State
University, Columbus, OH, USA

Changjun Liu (86)
Department of Flood and Drought Disaster
Reduction, China Institute of Water Resources
and Hydropower Research, Beijing, China


Huixin Liu (48)
The Bartlett Centre for Advanced Spatial
Analysis, University College London, UK

Martin Loidl  (83)
Department of Geoinformatics, University of
Salzburg, Austria


Nik Lomax  (21)
School of Geography, University of Leeds, UK

Jed A. Long  (49)
Western University, London, Canada


Maryam Lotfian  (8)
Institute INSIT, School of Business and
Engineering Vaud, University of Applied
Sciences and Arts Western Switzerland,
Yverdon-les-Bains, Switzerland

Binbin Lu  (39, 40)
School of Remote Sensing and Information
Engineering, Wuhan University, Hubei, China


Peter M. Macharia (79)
Department of Public Health, Institute of
Tropical Medicine, Antwerp, Belgium;
Population & Health Impact Surveillance Group,
Kenya Medical Research Institute-Wellcome
Trust Research Programme, Nairobi, Kenya;
Centre for Health Informatics, Computing, and
Statistics, Lancaster Medical School, Lancaster
University, UK


Ivan Majic  (63, 82)
Institute of Geodesy, Graz University of
Technology, Austria


Prestige Tatenda Makanga (79)
Surveying and Geomatics Department, Faculty
of Science and Technology, Midlands State
University, Gweru, Zimbabwe; Climate and
Health Division, Centre for Sexual Health and
HIV/AIDS Research, Zimbabwe


Milad Malekzadeh  (49)
Western University, London, Canada

Huy Quang Man (46)
VNU University of Science, Hanoi, Vietnam


Ed Manley  (45)
School of Geography, University of Leeds, UK

Robert Manning Smith  (50)
The Bartlett Centre for Advanced Spatial
Analysis, University College London, UK


Henry Martin  (11)
Institute of Cartography and Geoinformation,
ETH Zürich, Switzerland


Kazushi Matsuo  (51)
University of Tsukuba, Japan

Grant McKenzie  (9, 16, 61)
Spatial Analysis Lab, McGill University,
Montréal, Canada

Bouhadjar Meguenni  (53)
Spatial Reference Information Systems
Department, Space Techniques Center, Oran,
Algeria

Wouter Meulemans  (10)
TU Eindhoven, The Netherlands

Alexander Michels  (52)
CyberGIS Center for Advanced Digital and
Spatial Studies, University of Illinois
Urbana-Champaign, IL, USA

Jennifer A. Miller  (77)
Department of Geography and the Environment,
University of Texas at Austin, TX, USA


- Salim Miloudi  (53)
Spatial Reference Information Systems
Department, Space Techniques Center, Oran,
Algeria
- Richard Milton  (38)
Bartlett Centre for Advanced Spatial Analysis,
University College London, UK
- Franz-Benjamin Mocnik  (54)
University of Twente, Enschede, The
Netherlands; Paris Lodron University of
Salzburg, Austria
- Alan T. Murray  (4)
Department of Geography, University of
California at Santa Barbara, CA, USA
- Mortimer M. Müller  (62)
Institute of Silviculture, University of Natural
Resources and Life Sciences, Vienna, Austria
- Alicja Najwer  (42)
Adam Mickiewicz University, Poznan, Poland
- Maurizio Napolitano (75)
Fondazione Bruno Kessler, Trento, Italy
- Arpit Narechania  (55)
Georgia Institute of Technology, Atlanta, GA,
USA
- Nico Neureiter  (71)
Department of Geography, University of Zurich,
Switzerland; NCCR Evolving Language,
University of Zurich, Switzerland
- Linh Xuan Nguyen (46)
VNU University of Science, Hanoi, Vietnam
- Andreas Niekler  (14)
Computational Humanities, Leipzig University,
Germany
- Hayato Nishi  (56)
Graduate School of Social Data Science,
Hitotsubashi University, Tokyo, Japan
- Evgeny Noi  (57)
Department of Geography, University of
California Santa Barbara, CA, USA
- Martin Nöllenburg  (2)
Algorithms and Complexity Group, TU Wien,
Austria
- Olakunmi Ogunyemi (79)
Lagos State Ministry of Health, Nigeria
- Abimbola Olaniran (79)
Royal Tropical Institute, Amsterdam, The
Netherlands
- Tope Olubodun (79)
Department of Community Medicine and,
Primary Care, Federal Medical Centre ,
Abeokuta, Abeokuta, Ogun, Nigeria
- Itohan Osayande (79)
School of Human Sciences, University of
Greenwich, London, UK
- Frank O. Ostermann  (66)
Faculty of Geo-Information Science and Earth
Observation (ITC), University of Twente,
Enschede, The Netherlands
- Mustafa Can Ozkan  (58)
SpaceTimeLab, University College London, UK
- Jinwoo Park  (52)
CyberGIS Center for Advanced Digital and
Spatial Studies, University of Illinois
Urbana-Champaign, IL, USA
- Zhan Peng  (59)
Graduate School of Information Sciences,
Tohoku University, Sendai, Japan
- Lukas Plätz (18)
Ruhr University Bochum, Germany
- J. Gary Polhill  (38)
The James Hutton Institute, Aberdeen, UK
- Nina Polous  (60)
Institute of Geography and Regional Science,
University of Graz, Austria
- Niki Popper (63)
dwh GmbH, Vienna, Austria; Institute of
Information Systems Engineering, TU Vienna,
Austria
- Manon Prédhumeau  (45)
School of Geography, University of Leeds, UK
- Dan Qiang  (61)
Platial Analysis Lab, Department of Geography,
McGill University, Montréal, Canada
- Peter Ranacher  (71)
URPP Language and Space, University of
Zurich, Switzerland; Department of Geography,
University of Zurich, Switzerland; NCCR
Evolving Language, University of Zurich,
Switzerland
- Martin Raubal  (84)
Institute of Cartography and Geoinformation,
ETH Zürich, Switzerland


- Tumasch Reichenbacher  (95)
Department of Geography, University of Zurich,
Switzerland
- Liliang Ren (86)
State Key Laboratory of Hydrology-Water
Resources and Hydraulic Engineering, College of
Hydrology and Water Resources, Hohai
University, Nanjing, China
- Darja Reuschke  (49)
University of Southampton, UK
- Caitlin Robinson  (20)
School of Geographical Sciences, University of
Bristol, UK
- Amira Roess  (76)
Department of Global and Community Health,
George Mason University, Fairfax, VA, USA
- Carolina Rojas  (69)
Pontificia Universidad Católica de Chile,
Santiago, Chile; CEDEUS, Santiago, Chile
- Farnoosh Roozkhosh (70)
Department of Geography, University of Georgia,
Athens, GA, USA
- Robert Roth  (93)
Cartography Lab, Department of Geography,
University of Wisconsin-Madison, WI, USA
- Dominik Rothschedl (63)
dwh GmbH, Vienna, Austria
- David Röbl  (62)
Institute of Geodesy, Graz University of
Technology, Graz, Austria
- Doug Salt  (38)
The James Hutton Institute, Aberdeen, UK
- Naratip Santitissadeekorn  (36)
Department of Mathematics and Physics,
University of Surrey, UK
- Victoria Scherelis  (65)
Zurich University of Applied Sciences,
Wädenswil, Switzerland; University of Zurich,
Switzerland
- Joris Y. Scholl (5)
Ruhr-Universität Bochum, Germany
- Johannes Scholz  (62, 63, 82)
Institute of Geodesy, Graz University of
Technology, Graz, Austria
- Nadia Shafaeipour  (66)
Faculty of Geo-Information Science and Earth
Observation (ITC), University of Twente,
Enschede, The Netherlands
- Urmi Shah (67)
School of Computing & Mathematical Sciences,
University of Greenwich, UK; GeoLytx, London,
UK
- Arun Sharma (3)
Department of Computer Science & Engineering,
University of Minnesota, Minneapolis, MN, USA
- Shashi Shekhar (3)
Department of Computer Science & Engineering,
University of Minnesota, Minneapolis, MN, USA
- Alex Singleton  (20, 87)
Department of Geography and Planning,
University of Liverpool, UK
- Aidan Slingsby (68)
City, University of London, UK
- Nazia Ferdause Sodial (68)
City, University of London, UK
- Bettina Speckmann  (10)
TU Eindhoven, The Netherlands
- Stefan Steiniger  (69)
Pontificia Universidad Católica de Valparaíso,
Valparaíso, Chile; CEDEUS, Santiago, Chile
- John G. Stell  (37)
School of Computing, University of Leeds, UK
- Behnam Tahmasbi (70)
Dept of Civil and Environmental Engineering,
University of Maryland, College Park, MD, USA
- Takuya Takahashi  (71)
Department of Geography, University of Zurich,
Switzerland
- Garavig Tanaksaranond  (44)
Department of Survey Engineering, Faculty of
Engineering, Chulalongkorn University, Bangkok,
Thailand
- Faith Taylor  (94)
Department of Geography, King's College
London, UK; Centre for Advanced Spatial
Analysis, University College London, UK


- Jean-Claude Thill (90)
Department of Geography and Earth Sciences,
University of North Carolina at Charlotte, NC,
USA
- Richard Timmerman  (39, 40)
School of Geographical Sciences, University of
Bristol, UK
- Sabine Timpf  (72)
Geoinformatics Group, University of Augsburg,
Germany
- Guillaume Touya  (23, 73)
LASTIG, Univ Gustave Eiffel, IGN-ENSG,
F-77420 Champs-sur-Marne, France
- Binh Quoc Tran  (46)
VNU University of Science, Hanoi, Vietnam
- Ricardo Truffello  (69)
Pontificia Universidad Católica de Chile,
Santiago, Chile; CEDEUS, Santiago, Chile
- Morito Tsutsumi (51)
University of Tsukuba, Japan
- Joseph V. Tuccillo  (25, 74)
Oak Ridge National Laboratory, TN, USA
- Munazza Usmani  (75)
University of Trento, Italy; Fondazione Bruno
Kessler, Trento, Italy
- Harald Vacik  (62)
Institute of Silviculture, University of Natural
Resources and Life Sciences, Vienna, Austria
- Nathan van Beusekom  (10)
TU Eindhoven, The Netherlands
- Thomas C. van Dijk  (5)
Ruhr-Universität Bochum, Germany
- Maarten van Steen  (66)
Digital Society Institute (DSI), University of
Twente, Enschede, The Netherlands
- Priyanka Verma  (16)
Platial Analysis Lab, Department of Geography,
McGill University, Montréal, Canada
- Judith A. Verstegen  (4)
Department of Human Geography and Spatial
Planning, Utrecht University, The Netherlands
- Emma Von Hoene  (76)
Department of Geography and Geoinformation
Science, George Mason University, Fairfax, VA,
USA
- Boyu Wang  (81)
Department of Geography, University at Buffalo,
NY, USA
- Haoyu Wang  (77)
Department of Geography and the Environment,
University of Texas at Austin, TX, USA
- Jia Wang  (67, 79)
School of Computing & Mathematical Sciences,
University of Greenwich, UK
- Meihui Wang  (44)
Department of Civil, Environmental and
Geomatic Engineering, University College
London, UK
- Shaowen Wang  (52)
CyberGIS Center for Advanced Digital and
Spatial Studies, University of Illinois
Urbana-Champaign, IL, USA
- Xinglei Wang  (80)
SpaceTimeLab for Big Data Analytics,
University College London, UK
- Yiyu Wang  (78)
School of Geography, University of Leeds, UK
- Matthias Wastian (63)
dwh GmbH, Vienna, Austria
- Robert Weibel  (30)
Department of Geography, University of Zürich,
Switzerland
- Franz Welscher  (63, 82)
Institute of Geodesy, Graz University of
Technology, Austria
- Christian Werner  (83)
Department of Geoinformatics, University of
Salzburg, Austria
- Martin Werner (7)
Technical University of Munich, Germany
- René Westerholt  (11)
Department of Spatial Planning, TU Dortmund
University, Germany
- Nina Wiedemann  (11, 84)
Institute of Cartography and Geoinformation,
ETH Zürich, Switzerland
- Daniel Wiegrefe  (14)
Image and Signal Processing Group, Leipzig
University, Germany


- Jan Winkler  (30)
Department of Environmental Systems Science,
Swiss Federal Institute of Technology, Zürich,
Switzerland
- Sarah Wise  (48, 50)
The Bartlett Centre for Advanced Spatial
Analysis, University College London, UK
- Kerry L. M. Wong (79)
Faculty of Epidemiology and Population Health,
London School of Hygiene and Tropical
Medicine, UK
- Jo Wood  (10)
City, University of London, UK
- Carson Woody  (85)
Human Geography Group, Oak Ridge National
Laboratory, TN, USA
- Jules Wolms  (2)
Algorithms and Complexity Group, TU Wien,
Austria
- Guohui Xiao  (28)
Department of Information Science and Media
Studies, University of Bergen, Norway
- Ningchuan Xiao  (47)
Department of Geography, The Ohio State
University, Columbus, OH, USA
- Rongbo Xu (19)
Centre for Advanced Spatial Analysis,
University College London, UK
- Ikuho Yamada  (56)
Center for Spatial Information Science, The
University of Tokyo, Japan
- Jing Yao  (86)
Urban Big Data Centre, School of Social and
Political Sciences, University of Glasgow, UK
- X. Angela Yao (70)
Department of Geography, University of Georgia,
Athens, GA, USA
- Zi Ye  (87)
Department of Geography and Planning,
University of Liverpool, UK
- Marcus A. Young  (88)
Transportation Research Group, University of
Southampton, UK
- Jingyan Yu  (36)
Institute of Geography and Sustainability (IGD),
Faculty of Geosciences and Environment,
University of Lausanne, Switzerland
- May Yuan  (89)
Geospatial Information Sciences, The University
of Texas at Dallas, TX, USA
- Zhendong Yuan (7)
Utrecht University, The Netherlands
- Alexander Zaft (5)
Universität Würzburg, Germany
- Zichao Zeng  (44)
Department of Civil, Environmental and
Geomatic Engineering, University College
London, UK
- Han Zhang (90)
School of Urban Planning and Design, Peking
University Shenzhen Graduate School, Shenzhen,
Guangdong, China; Key Laboratory of Earth
Surface System and Human-Earth Relations of
Ministry of Natural Resources of China, Peking
University Shenzhen Graduate School, Shenzhen,
Guangdong, China
- Hongyu Zhang  (9)
Platial Analysis Lab, McGill University,
Montréal, Canada
- Jing Zhang (91)
School of Geographical Sciences, University of
Bristol, UK
- Qianheng Zhang (93)
HGIS Lab, Department of Geography,
University of Washington, Seattle, WA, USA
- Wenlan Zhang  (94)
Centre for Advanced Spatial Analysis,
University College London, UK
- Xianghui Zhang  (80, 92)
SpaceTimeLab for Big Data Analytics,
University College London, UK
- Xiaoxiang Zhang (86)
Department of Geographic Information Science,
College of Hydrology and Water Resources,
Hohai University, Nanjing, China
- Pengjun Zhao (27)
School of Urban Planning and Design, Peking
University Shenzhen Graduate School, Shenzhen,
Guangdong, China; Key Laboratory of Earth
Surface System and Human-Earth Relations of
Ministry of Natural Resources of China, Peking
University Shenzhen Graduate School, Shenzhen,
Guangdong, China; College of Urban and
Environmental Sciences, Peking University,
Beijing, China

Qunshan Zhao (29)
Urban Big Data Centre, School of Social and
Political Sciences, University of Glasgow, UK

Chen Zhong  (94)
Centre for Advanced Spatial Analysis,
University College London, UK


Zhiyong Zhou  (30)
Department of Geography, University of Zürich,
Switzerland

Dingyi Zhuang  (33)
Department of Urban Studies and Planning,
Massachusetts Institute of Technology (MIT),
Cambridge, MA, USA

Donatella Zingaro  (95)
Department of Geography, University of Zurich,
Switzerland

Alexander Zipf (7)
GIScience Chair, Heidelberg University,
Germany

Zbigniew Zwoliński  (42)
Adam Mickiewicz University, Poznan, Poland

Seda Şalap-Ayça  (64)
Department of Earth, Environmental, and
Planetary Sciences, Brown University,
Providence, RI, USA; Institute at Brown for
Environment and Society, Brown University,
Providence, RI, USA; Department of Earth,
Geographic, and Climate Sciences, University of
Massachusetts, Amherst, MA, USA

Do You Need Instructions Again? Predicting Wayfinding Instruction Demand

Negar Alinaghi  

Geoinformation, TU Wien, Austria

Tiffany C. K. Kwok  

Institute of Cartography and Geoinformation, ETH Zürich, Switzerland

Peter Kiefer  

Institute of Cartography and Geoinformation, ETH Zürich, Switzerland

Ioannis Giannopoulos  

Geoinformation, TU Wien, Austria

Institute of Advanced Research in Artificial Intelligence (IARAI), Austria

Abstract

The demand for instructions during wayfinding, defined as the frequency of requesting instructions for each decision point, can be considered as an important indicator of the internal cognitive processes during wayfinding. This demand can be a consequence of the mental state of feeling lost, being uncertain, mind wandering, having difficulty following the route, etc. Therefore, it can be of great importance for theoretical cognitive studies on human perception of the environment. From an application perspective, this demand can be used as a measure of the effectiveness of the navigation assistance system. It is therefore worthwhile to be able to predict this demand and also to know what factors trigger it. This paper takes a step in this direction by reporting a successful prediction of instruction demand (accuracy of 78.4%) in a real-world wayfinding experiment with 45 participants, and interpreting the environmental, user, instructional, and gaze-related features that caused it.

2012 ACM Subject Classification Computing methodologies → Activity recognition and understanding; Computing methodologies → Supervised learning by classification

Keywords and phrases Wayfinding, Navigation Instructions, Urban Computing, Gaze Analysis

Digital Object Identifier 10.4230/LIPICs.GIScience.2023.1

Supplementary Material *Dataset:* <https://geoinfo.geo.tuwien.ac.at/resources/>

Acknowledgements We would like to thank our colleagues from Vienna University of Technology, Dr. Markus Kattenbeck, for suggesting the use of the Big Five Personality Traits test, and Antonia Golab for collecting the valuable data used for this work.

1 Introduction

Human-computer interaction (HCI) in wayfinding and pedestrian navigation has attracted much attention in recent years [27]. Reducing cognitive load in a complex task such as wayfinding is an important goal in this domain. Efforts in this area range from working on the structure, content, and presentation of navigation information to better adapting it to the needs of users. The current research trend in navigation assistance systems provides instructions in various modalities, from conventional turn-by-turn instructions with map visualization to auditory instructions based on landmarks and, more recently, visual instructions with augmented reality (e.g., [44, 10]). The performance of these different modalities is evaluated using various metrics including travel time, number of errors, deviation from shortest/fastest path, cognitive demand, subjective ratings, etc. (see Section 2). One possible indicator which is less explored is how often the user asks for the instruction after first receiving it, i.e., the instruction demand. The behavior of requesting more information



© Negar Alinaghi, Tiffany C. K. Kwok, Peter Kiefer, and Ioannis Giannopoulos; licensed under Creative Commons License CC-BY 4.0

12th International Conference on Geographic Information Science (GIScience 2023).

Editors: Roger Beecham, Jed A. Long, Dianna Smith, Qunshan Zhao, and Sarah Wise; Article No. 1; pp. 1:1–1:16

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1:2 Do You Need Instructions Again?

or the same information again can also be considered as an indicator of how cognitively demanding the processing of this information is. However, we still do not know whether such behavior is caused by the complexity of the environment, the user's personal characteristics or spatial abilities, the content of the instructions and how they are conveyed, or a combination of all these factors.

Being able to predict the overall demand for instructions shortly after receiving a navigation instruction, can be very useful in developing more customized navigation aids. On the other hand, knowing why people need more instructions and what factors trigger this demand in users is of great importance for cognitive studies on human perception of the environment. For instance, the need for repetition of navigation instructions can be seen as a prototypical behavior of feeling lost or needing reassurance. Therefore, predicting this need can as well help us predict these cognitive states during wayfinding. Theoretical studies and empirical evidence suggest that cognitive demand in wayfinding is strongly influenced by user-, environment-, and assistance-related factors [39, 15]. Requesting instructions is one of the many activities performed during this process and can be considered a proxy for cognitive demand. Machine learning (ML) has shown promising results, not only in predicting such activities but also in partially explaining the predictions. The latter is crucial when it comes to interpreting the results and extending knowledge about the causality of actions, here e.g., instructional needs.

In this paper, we show that instruction demand in a pedestrian wayfinding experiment can be predicted with reasonably high accuracy using ML techniques. This prediction is not a black box; rather, our results suggest that it can be interpreted by environmental features, user- and instruction-related features, and gaze features. These findings are the result of an outdoor wayfinding experiment conducted in the city with 45 participants navigating to two different known and unknown destinations using an audio-assisted landmark-based navigation system. Participants were allowed to ask for auditory instructions at any time and as often as they wished. While walking, their behavioral data in the form of eye movements and trajectory were recorded by eye-tracking glasses and a high-precision GNSS antenna (see Section 3.1).

We trained several classifiers, namely Support Vector Machines (SVM), RandomForest, and XGBoost on a variety of gaze features, environmental features, user- and instruction-related features. Our analysis shows that instruction demand can be predicted shortly after the first instruction (within two seconds), mainly based on the complexity of the environment and user characteristics. Through several experimental setups, we found that a minimal subset of 21 features (a combination of the above factors) leads to an accuracy of 78.9% on unseen data, making our prediction approach beneficial for real-time applications and giving us a better understanding of why people may need more informative assistance in wayfinding.

2 Related Work

With the goal of predicting instruction needs in an auditory-aided wayfinding task, we examined the existing literature from several perspectives: the evaluation and perception of navigation instructions, the processing of instructional information using gaze analysis, and exploring the use of machine learning in predicting wayfinding activities and states.

2.1 Research on Navigational Instructions

Efforts have been made to optimize wayfinding support systems through the structure of instructions [25, 41, 35], use of landmarks [34, 31, 12], and modality of information presentation [20, 10]. One of the first papers to address the conceptualization of route instructions is by Klippel et al. [24], who proposed a set of wayfinding choremes as mental conceptualizations of route guidance elements used to simplify visualization of turn-by-turn information. Another method to reduce instruction complexity is the Spatial Chunking method, where unnecessary instructions are chunked together to reduce complexity [23]. A second aspect besides simplifying instructions is how much spatial information they convey. Krukar et al. [26], addressed this matter by introducing orientation instructions that combine local and global route information. In a study with 84 participants, they evaluated the performance of these instructions by measuring the memorability of the instructions and showed that they conveyed survey information without interfering with the retrieval of route information. A systematic review of navigation systems for people with dementia was conducted by Pillette et al. [32] to compare common evaluation standards. They reviewed 23 papers, including indoor, outdoor, and VR-based experimental designs, in terms of presentation modality, navigation content, and timing of presentation. Most objective measures introduced for evaluation were the number of errors, time to complete the navigation task, arrival at destination, and the number of times participants asked questions or received outside assistance. Most subjective measures were obtained either by experimenters observing participants' behavior, hesitation, or difficulty in completing the task, or by interviews and questionnaires.

The work most closely related to ours is that of Golab et al. [17]. The authors used survival analysis to model the times at which participants needed navigation instruction, accounting for personal, environmental, and route-related variables. They reported that “participants request a route instruction later as a function of their age, on segments longer than 120m and in unfamiliar conditions if they score below average on the personality trait extraversion.” This is initial evidence from a real-world wayfinding experiment, reporting user-related and environmental aspects to be influential on the timing of the instructions.

2.2 Gaze Analysis in Instruction-based Wayfinding

Eye tracking provides insight into human cognitive processes [18, 11] complementing standard behavioral responses. Previous work has shown the potential of gaze for user modeling, such as including user's attention [1], cognitive load assessment [6], and spatial decision making [40, 22]. The knowledge obtained from user modeling can further be used for adapting the system behavior [16]. An example of gaze analysis in wayfinding is Brügger's study [8] where they examined the impact of navigation system behavior on human navigation behavior and performance in an outdoor experiment with 64 participants. They measured cognitive function and scene complexity using fixation frequency and duration, finding that average fixation duration varied across different tasks, i.e., incidental knowledge acquisition and knowledge retrieval.

De Cock et al. [10] used gaze behavior analysis to investigate the nature of navigation instructions in an indoor experiment with a VR-based adaptive route guidance system. They focused on photos or icons at start and end points, and photos or 3D simulations at turn junctions. They mainly used dwell time for their analysis and found that the detailed information of a static photo instruction is more difficult to transfer to the environment. Very recently in a paper by Ludwig et al. [29], the instruction needs in a real-world indoor

multi-level wayfinding experiment were analyzed using the normalized mean square error between the observed dwell time distribution and its estimation from the distribution of the aggregated fixations between two routing instructions which were generated automatically by their system incorporating the most salient landmark close to the user. Their result suggested that instruction need tends to increase when there is a change in direction or level. These papers show practical evidence that gaze analysis can reveal aspects of the cognitive demands of processing navigation instructions.

2.3 Machine Learning for Wayfinding Activities' Prediction

The use of ML techniques for prediction and classification tasks in wayfinding has become increasingly important in recent years. Alinaghi et al. [4] predicted the direction in which the wayfinder would like to turn a few seconds before the turn action is performed, based on gaze behavior and environmental complexity. They tested several ML techniques, including SVM, DecisionTree, and XGBoost, and reported that XGBoost outperformed the other two with 91% accuracy. In another paper [3], the authors analyzed the effect of familiarity with the environment using the pre-trained XGBoost model from their earlier work and were able to show that the gaze behavior of familiar and unfamiliar wayfinders differed as they approached a turn decision point. There, the authors introduced a terminology as *matching-to-action* phase of wayfinding, which in their context refers to the part of the route from the point where an instruction is given to the turn to which the instruction refers. We have segmented our trial routes based on this notion (see Section 3.2). Liao et al. [28] trained a RandomForest classifier with gaze features from 38 participants in a real-world outdoor study to classify five common wayfinding tasks: Self-location and orientation, target search in the local environment, target search on the map, route memorization, and walking to the target, with an overall accuracy of 67%.

Another related experience with the RandomForest classifier is reported by Zhu et al. [44]. In a VR-based indoor study with 30 participants, the authors collected EEG recordings from participants performing a series of 10 wayfinding trials of varying difficulty, each exploring a portion of the virtual reality model (i.e., different sets of origin and destination locations were defined at different distances and on multiple floors for each trial). Using a combination of objective measurements (e.g., frequency of inputs to the VR controller) and behavioral recordings from two independent observers, a classifier was trained to predict uncertainty time segments during navigation trials. The overall predictive power of the model was reported to be 0.70 as measured by the area under the Receiver Operating Characteristics Curve (ROC-AUC). Although most work on activity recognition in the wayfinding domain uses the RandomForest model due to its simplicity in training and explainability, higher performance with other models is also reported in the broader domain of human activity recognition. For example, XGBoost, as an ensemble model of tree-based architectures that can better model more complex relationships and is as explainable as other tree-based models, has been successfully used in the literature for many different tasks (see, e.g., [5, 43]).

3 Data Collection, Pre-processing and Feature Extraction

This section summarizes the details of the data collection procedure, the pre-processing steps, and finally the feature extraction methods. The data we report on here was collected in 2020 and was first described in [17]¹. Parts of the data have already been used for analyses

¹ Parts of the data used in the current paper, will be made available at: <https://geoinfo.geo.tuwien.ac.at/resources/> (DOI: 10.5281/zenodo.4298703).

published in [4], [3] and [2]. The original dataset has 104 trials recorded from 52 participants (27 female and 25 males, $M(\text{age}) = 26$ years, $SD(\text{age}) = 8.3$). However, due to some sensor malfunctioning and data loss in eye-tracking data, here we have analyzed 71 trials from 45 participants.

3.1 Data Collection

The study was a within-subject experiment with two phases: an online phase for registration, demographic, Big Five personality traits [33], and the Spatial Strategies Questionnaire (FRS) [30] data collection; and an in-situ phase for recording participants' eye movements (using PupilLabs Invisible glasses with 200 Hz recording frequency) and trajectory data (using a PPM 10-xx38 GNSS receiver) as they walked familiar and unfamiliar routes². The familiar trial was conducted in a region and to a destination that the participant reported being completely familiar with, as opposed to the unfamiliar trial. In both cases, however, participants were asked to walk a pre-calculated route by following the auditory turn-by-turn, German-language, landmark-based navigation instructions³ provided to them upon request. Participants were given a clicking device that they held in their hands and could click on when and as often as they wanted. The obligation to follow the predefined route (which was unknown to participants) prompted all participants to request navigation instructions to find their way, whether or not they were in the familiar condition. This design provided us with the opportunity to examine primarily *when* they need instructions and then *why* they need the same instruction more than once.

3.2 Data Pre-processing

To determine the motives behind instruction demand, we processed four of the data sources: GPS tracks to obtain the environmental features, online-phase data to obtain the demographics, personality traits, and spatial strategy scales, eye-tracking records for gaze behavior, and navigation instructions for their length and content. Of these, we only needed to preprocess the GPS tracks to match them with the Open Street Maps (OSM) data and extract urban complexity measures from them. Figure 1 depicts the six steps required for this preprocessing. First, we cleaned the GPS data and smoothed it by preserving the timestamps. Then, using the intersection framework [14], we extracted street intersections from the OSM data and matched them to the GPS tracks. The instruction-request events were also matched to the GPS tracks based on their timestamps. Then, we segmented the route based on the idea of *matching-to-action phase* [3] (See Subsection 2.3) with the small difference that we do not start this segment with the instruction request event, but with the previous turn intersection to also track how far the request event is from the last turn decision point, and we call the outcome chunks of the route "*unified segments*". We also call the part of the route between any two intersections (whether it is a turn or not) a "*segment*". Finally, a buffer of 30m (large enough to cover the buildings and Points of Interest (PoIs) from both sides of the road) was considered around the unified segments to extract building footprints and their relative attributes.

As for the prediction class, we labeled each event of the first request as "*1-click*", "*2-clicks*", or "*more-clicks*". The 1-click class means that the instruction was requested only once, while the 2-clicks and more-clicks classes mean that the instruction was requested two or more

² Participants indicated their familiarity in three levels (region, route, and landmark) in the online phase.

³ Route instruction pattern: TURN LEFT [IMPERATIVE] AT CAFE FABRIK [LANDMARK].

times. For example, in Figure 1, E1 and E2 events are marked as instances of class “1-click”, E3 as an instance of class “more-clicks”, and E4 as an instance of class “2-clicks”. In total, we had 234 samples of such first-request events in the unified segments, with an unbalanced distribution across the classes (1-click = 68.803%, 2-clicks = 20.940%, and more-clicks = 10.256%).



■ **Figure 1** Preprocessing steps applied to the GPS tracks: smoothing (green line), OSM junction extraction and mapping to the route (black and yellow circles), instruction-event alignment (red circles), segmentation into unified segments with class labels (E1–E4), and extraction of land use and POI information from OSM in a 30m buffer around the route.

3.3 Feature Extraction

Selecting the right features for training a machine learning model is a very important step. Of course, any algorithm trained on any set of features will yield a model, but according to the principle of “garbage-in, garbage-out”, the model trained on inappropriate features, even if it performs well on the training data, fails to generalize and thus explain the results. Following the model of wayfinding decision situations [15], we selected features that reflect the complexity of the environment and instructions, as well as the characteristics of the user.

We extracted four groups of features from our data sources: 41 *Environmental*, 16 *Instruction-*, 12 *User-related*, and six *Gaze* features, yielding a total of 75 features. Table 1 summarizes these features into subcategories. For the environment category, we extracted seven features for the segments, including the length of the route and road segments, the elapsed time since the start of the trial, the distance to the previous and next non-turn intersections, and the distance to the previous and next turn intersections. The elapsed time, while not a direct environmental feature, may serve as a proxy for length/speed and may also be related to working memory (see Section 6). Based on the landcover codes of Urban Atlas data⁴, we extracted 13 features describing land use in the buffer around the unified segments. These features are in fact the proportion of buffer area for each land use. For example, we calculated how much of the area is occupied by “*Green-urban-areas*” or “*Sports-and-leisure-facilities*”. Using the same approach, we extracted from the OSM data the semantic label of POIs⁵ and their density along the unified segments. For example, by counting the number of POIs with the amenity tag “*shop*” per unit area along the unified segments, we calculated the density of the POI “*shop*”. In total, there were 10 amenities in our experimental regions and we calculated the density for each. We also calculated the total POI density to have a measure of the visual complexity of the environment. This yields a total of 21 POI-related features for the environment category.

⁴ <https://land.copernicus.eu/user-corner/technical-library/urban-atlas-mapping-guide/view>

⁵ <https://wiki.openstreetmap.org/wiki/Key:amenity>

■ **Table 1** Extracted features for predicting instruction demand: POIs (environmental and instruction-related) are extracted from OSM with standard amenity types (e.g. shop, touristic, etc.).

Environmental Features	41 (7+13+21)	Instruction Features	16 (2+14)
<i>unified-segment</i>	7	<i>length-related</i>	2
distance from/to previous/next turn junctions	2	number of words	1
distance from/to previous/next non-turn junctions	2	number of characters	1
segment-length	1	<i>content-related</i>	14
route-length	1	OSM PoI	11
time passed since start	1	landmark OSM type	1
<i>landuse</i>	13	contains-street-names (boolean)	1
<i>PoI</i>	21	last instruction (boolean)	1
User Features	12 (3+5+4)	Gaze Features	6
<i>demographics</i>	3	fixation count	1
gender (binary)	1	min/max/sd fixation	3
age (in years)	1	mean fixation duration	1
familiarity (binary)	1	fixation duration skewness	1
<i>Big Five Personality traits</i>	5		
<i>Spatial Strategies Questionnaire FRS</i>	4		

For instructions, since we assumed that the length of the instruction may affect the demand, we calculated the number of words and characters in the instructions. We also extracted 11 OSM-PoI features as one-hot encoding (the 11 PoIs used as landmarks in the instructions), and three features describing whether the instruction contains a street name, whether it is the last instruction, or what kind of spatial object it refers to (e.g., point or an area). In terms of user-related characteristics, we had age, gender, and a binary measure of familiarity as demographic data from the online study; five values for personality traits (openness, conscientiousness, extraversion, agreeableness, and neuroticism) obtained from the Big Five personality test; and four values for spatial strategy scales (preference for egocentric, allocentric, cardinal directions, and the sense of direction score).

Finally, for the gaze-based features, as one goal of our analysis was also to determine how quickly we could predict instruction demand after the first request, we segmented the gaze data into 1- to 10-second windows immediately after the first event, and extracted features within these windows. In this way, we were able to find the minimal set of gaze data for the prediction task. Eye-tracking datasets were collected in a mobile eye-tracking scenario with free head/body movements. It is well known that head movements strongly influence the calculation of saccade length and velocity [2], so no saccadic features were computed. We, therefore, computed only basic fixation-based features. Fixations were extracted using the Dispersion-Threshold Identification (I-DT) algorithm [36] (gaze-dispersion threshold: 1 deg; duration threshold: 100 ms). Fixation count and five statistical measures from fixation duration (mean, minimum, maximum, standard deviation, and skewness of the frequency distribution), were extracted from the data.

By extracting all these features, we obtained a dataset of size $234 * 75$. These 75 features were extracted based on our assumptions about influential factors on instruction demand. After training the models and analyzing the feature importance (see Section 5), we were able to prune this list and extract the most relevant features and interpret their effect.

4 Machine Learning Experiments

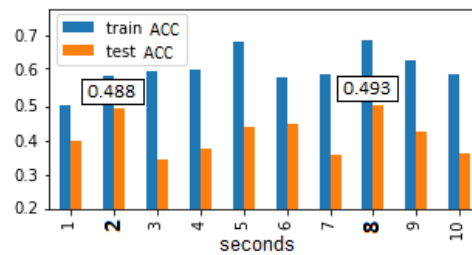
To predict instruction demand, i.e., how often a wayfinder requests an instruction, we trained three classifiers: SVM, RandomForests, and XGBoost. The choice of models was based on previous experience [4] and other successful reports of the performance of these models in the literature of both human activity recognition and wayfinding (see Section 2.3). For instance, our experience with the SVM-RBF, CART, Random Forest, and XGBoost algorithms for a similar task of predicting pedestrians' turning activity based on their gaze behavior (whether they turn left/right at an intersection or continue straight ahead) yielded test accuracies of $.58 \pm .05$, $.61 \pm .06$, $.77 \pm .06$, and $.91 \pm .08$, respectively. Before applying any of these models, we first normalized the data and converted all categorical values to numerical values. To deal with the imbalance of the data, we also compared two methods, sample weighting, and oversampling. It is known that if the number of samples in the minority class is too small and the samples are much farther away from the other classes, the highly weighted samples, although drawing the edge of the decision function to themselves, may not be effective enough to actually lie within the decision function [19]. However, we tried both methods and obtained better results ($\approx 5.6\%$) with Synthetic Minority Over-sampling Technique (SMOTE) [9], which may also be an observation of the same phenomenon in sample weighting. To find out how fast we can predict the demand, in a pilot testing, we trained an XGBoost classifier for each of the 1- to 10- second windows gaze data and plotted the performance metrics to select the optimal window size (see Section 5 for details).

Once we had the data ready, we set up the experimental pipeline in a way to avoid both data leaks and unwanted effects of participants' individuality in training and testing. We split the data into 70% and 30% (train, test) using the leave-one-group-out (LOGO) method to ensure that test participant samples were not part of the training. Then, oversampling was applied as a pipeline separately on the test data and within the 10-fold cross-validation to ensure that there was no data leakage not only between the training and test datasets but also within the validation folds. The training results were checked for overfitting by *mlogloss* and *merror* with cross-entropy loss function. Finally in order to interpret the results and prune the features, and see how much the model's predictions are influenced by every feature, we applied both the Tree SHapley Additive exPlanations (SHAP) method and feature importance by permutation (i.e., leaving the features one by one out ordered by their importance and monitor the drop in accuracy). The interpretation of the results using features' importance is presented and discussed in Section 6.

5 Results

This section summarizes the results of our analysis. First, we report on the SVM and RandomForest classifiers, which achieved test accuracy of 62.88% and 69.78%, respectively, when trained on all the data, being the highest accuracy in both cases. However, the XGBoost classifier outperformed both by an average of 11.2% across all feature combinations. Therefore, only the XGBoost results are presented here in more detail.

We begin with the gaze window sizes. Figure 2 plots the train and test accuracies for 1- to 10-second windows, with the 8- and 2-second windows representing the best and second-best prediction results for instruction demand. The fact that we can predict demand 6 seconds faster while losing less than 1% in accuracy makes window size two more attractive for the application domain. However, since the 8-second window still has the highest accuracy, we continue to report results for this window size.



■ **Figure 2** Compares the XGBoost results trained on the gaze data of different window sizes.

As explained in Subsection 3.3, we extracted different categories of features to determine which categories are most important for our prediction task. We conducted 15 experiments in which the pipeline was set to different categories of features or combinations thereof. Table 2 summarizes these results in terms of accuracy, f1 score, precision, recall, and Cohen’s kappa. Exp. 1, 2, 3, and 4 are based on single categories with 6, 16, 12, and 41, gaze, instruction, user, and environmental features, respectively. We also run some experiments with the subcategories of these features, which are summarized in Table 1. These experiments are labeled *.1 through *.3. For example, in Exp. 3.2, the model is trained with the five features of the Big Five personality test to see how personality alone can define instruction demand. In Exp. 6, the model was trained with 16 features that were among the top 4 features from Exp. 1, 2, 3, and 4. Exp. 7 is the result of applying the feature selection by permutation approach. In each iteration of the permutation step, a subset of the features was selected by removing each individual feature based on its importance rank and the model was trained on that subset. The best feature set to report was defined as the smallest subset with the least deviation from the best resulting model, in this case the model trained on all 75 features. As can be seen in Table 2, the three best performances of the model belong to experiments using all features (Exp. 5), permutation-selected features (Exp. 7), and a combination of the four most important features of individual categories (Exp. 6), with a test accuracy of 79.1%, 78.9%, and 77.4%, respectively. However, the kappa values show a better agreement result for the permutation-selected features. Using only gaze features, the model achieves the lowest performance of 49.3%. After that, the instruction-related features provide an accuracy of 58%. User-related features, despite having a smaller number of features than instruction-related features, provide the model with an accuracy of 65.6%. The environment-related features with the highest number of features among the individual categories are close to the best-performing model with less than 4% difference in test accuracy (i.e., 75.8%).

Figures 3 and 4 show the SHAP ranks of each feature category (i.e., Exp. 1, 2, 3, and 4) and the top three best-performing experiments (i.e., Exp. 5, 6, and 7), respectively. In Figure 3, the top four features from each category are used to train the model in Exp. 6. These features are distance to and from turn points and the nearest non-turn intersection, time elapsed since the start of the experiment, familiarity, sense of direction, openness, egocentric preference, length of the instruction, content of the instruction in terms of type of landmarks used and turn direction (left or right), number of fixations, standard deviation of fixations, average, and skewness of fixation duration. In Figure 4, environmental characteristics are among the top four important features in all cases (Exp. 5, 6, and 7). Among user-related features, which are the second most important category, gender, the personality trait of openness, and sense of direction are the most important ones. Instruction length, measured by the number of characters, and instruction content, measured by the type of landmark

1:10 Do You Need Instructions Again?

and turn direction, are as well present in the list of the most important features. Fixation features are at the bottom of all three lists, with average duration and maximum fixation being the most important ones. These results are further discussed in Section 6.

■ **Table 2** Summarizes the results of the trained XGBOOST classifier for different combinations of features in terms of accuracy, f1 score, precision, recall, and Cohen’s kappa. Experiment 7 with 21 features yields the best performance after experiment 5 with 75 features.

Exp.	Features	#	Split (LOGO)	Evaluation Metrics				
				Accuracy	F1 Score	Precision	Recall	Kappa
1	gaze	6	train	0.683	0.683	0.682	0.682	–
			test	0.493	0.493	0.495	0.443	0.432
2	instruction	16	train	0.689	0.690	0.692	0.689	–
			test	0.580	0.523	0.560	0.567	0.526
2.1	instruction length	2	train	0.513	0.506	0.511	0.512	–
			test	0.44	0.394	0.393	0.469	0.413
2.2	instruction content	14	train	0.586	0.565	0.627	0.608	–
			test	0.565	0.568	0.581	0.565	0.505
3	user	12	train	0.731	0.729	0.729	0.730	–
			test	0.656	0.632	0.678	0.641	0.632
3.1	user demographics	3	train	0.537	0.527	0.529	0.536	–
			test	0.527	0.485	0.492	0.584	0.415
3.2	user BigFive	5	train	0.682	0.683	0.689	0.683	–
			test	0.530	0.505	0.528	0.529	0.531
3.3	user FRS	4	train	0.634	0.634	0.634	0.634	–
			test	0.569	0.521	0.614	0.548	0.489
4	environment	41	train	0.855	0.854	0.855	0.854	–
			test	0.758	0.711	0.744	0.757	0.752
4.1	environment segment	7	train	0.841	0.836	0.849	0.840	–
			test	0.696	0.642	0.655	0.670	0.644
4.2	environment PoI	21	train	0.675	0.677	0.681	0.675	–
			test	0.656	0.638	0.702	0.682	0.638
4.3	environment landuse	13	train	0.668	0.671	0.680	0.668	–
			test	0.521	0.495	0.528	0.532	0.567
5	all	75	train	0.896	0.897	0.902	0.896	–
			test	0.791	0.746	0.758	0.757	0.742
6	manual selection of features based on their importance rank in experiments 1, 2, 3, and 4	16	train	0.879	0.879	0.882	0.879	–
			test	0.774	0.764	0.784	0.792	0.763
7	selection by permutation	21	train	0.896	0.896	0.897	0.896	–
			test	0.789	0.778	0.794	0.802	0.783

6 Discussion

Here we predicted the instruction demand in a pedestrian wayfinding scenario, after the first instruction was given, based on a set of feature categories. In a real-world application scenario, we assume that three of these categories, namely environmental features, user-related features, and instruction-related features, are fixed once the route to be navigated is selected. That is, once the navigation system computes the optimal route and generates the instructions for it, the environment- and instruction-related features can be easily computed. Similarly, user-related features can be collected in the sign-up information. In contrast, gaze behavior is not static and is heavily influenced by the task, stimulus processing, movements, and so on [11]. Gaze features should therefore be computed immediately after the first instruction. To determine how much gaze data should be recorded for this prediction, we tested different window sizes and found that as little as two seconds of fixation behavior recording can support the prediction with slightly lower accuracy than 8 seconds (**Exp. 1**). It is well known from the gaze analysis literature that fixation behavior can be interpreted in terms of frequency and duration as a sign of cognitive load on information processing, attention, and scene perception (see, e.g., [38, 21]). Our results are consistent with these findings, but also show that in a real-world situation, these features do not by themselves encode enough

Importance Rank	Gaze Features (Exp.1)	Instruction-related Features (Exp. 2)	User-related Features (Exp. 3)	Environmental Features (Exp. 4)
1	Gaze: fixation count	Inst: number of characters	User: familiarity	Env: distance to previous turn point
2	Gaze: average fixation duration	Inst: turn direction	User: FRS sense of direction	Env: time passed since start
3	Gaze: skewness of fixation duration	Inst: osm type	User: FRS egocentric	Env: distance to next intersection
4	Gaze: sd fixation	Inst: number of words	User: BigFive openness	Env: distance to the next turn point
5	Gaze: min fixation	Inst: last instruction	User: BigFive agreeableness	Env: poi shop
6	Gaze: max fixation	Inst: amenity	User: gender	Env: landuse green urban areas
7		Inst: shop	User: FRS cardinal	Env: distance to previous intersection
8		Inst: landmark- based	User: BigFive neuroticism	Env: poi leisure
9		Inst: contains street names	User: BigFive extraversion	Env: overall poi density
10		Inst: street	User: age	Env: landuse roads and associated lands
11		Inst: tourism	User: FRS allocentric	Env: route length
12		Inst: leisure	User: BigFive Conscientiousness	Env: landuse discontinous dense urban fabric
13		Inst: historic		Env: poi tourism
14		Inst: public-transport		Env: poi public-transport
15		Inst: natural		Env: segment length
16		Inst: man made		Env: landuse continous urban fabric
17				Env: poi density highway
18				Env: poi density natural
19				Env: poi density shop
20				Env: poi amenity

■ **Figure 3** The SHAP feature importance rank for Experiments 1 to 4. Top 4 features of each category is manually selected for training the model in experiment 6.

Importance Rank	All Features (Exp. 5)	Manually Selected Features (Exp. 6)	Permutation-selected Features (Exp. 7)
1	Env: distance to previous turn point	Env: distance to previous turn point	Env: distance to previous turn point
2	Env: time passed since start	Env: time passed since start	Env: time passed since start
3	Env: distance to the next turn point	Env: distance to next intersection	Env: distance to the next turn point
4	Env: distance to next intersection	Env: distance to the next turn point	Env: distance to next intersection
5	Env: segment length	Inst: osm type	Env: poi: shop
6	Env: poi shop	User: FRS egocentric	Env: distance to previous intersection
7	User: gender	Gaze: average fixation duration	User: FRS sense of direction
8	Env: poi leisure	User: familiarity	User: BigFive openness
9	User: BigFive openness	User: BigFive openness	Env: poi density amenity
10	Inst: amenity	Gaze: fixation count	User: gender
11	Env: poi highway	Inst: turn direction	Inst: number of characters
12	User: familiarity	Inst: number of characters	Inst: turn direction
13	Env: poi public-transport	Gaze: skewness of fixation duration	User: FRS allocentric
14	Env: distance to previous intersection	User: FRS sense of direction	Env: landuse green urban areas
15	Gaze: max fixation	Inst: number of words	Env: poi public-transport
16	Gaze: min fixation	Gaze: sd fixation	Gaze: average fixation duration
17	Env: poi density shop		User: FRS cardinal
18	Gaze: sd fixation		User: familiarity
19	Inst: turn direction		Gaze: min fixation
20	User: FRS sense of direction		Inst: osm type

■ **Figure 4** The SHAP feature importance rank of the three best-performing feature sets: All features, manually selected features, and permutation selected features. In all groups, environmental features are among the most important features.

information to predict instructional demand.

According to Table 2, instruction-related features (**Exp. 2**) are more informative for the model than gaze alone. Among them, the length of the instruction, encoded with the number of words as a language-independent proxy and the number of characters as a German-language proxy, is more important. This is consistent with the well-known word length effect, which states that longer words (which are common in German) are less well remembered [7]. We assume that both length measures correlate with the memory and recall performance of the instruction. This assumption is inline with previous findings suggesting that longer and more complex navigation instructions may lead to increased cognitive load and decreased wayfinding performance, making the instructions less memorable [23, 26]. The observation that landmark type and turn direction (two important content-related features) are also among the top five features, also supports this assumption. It means that *what information* and *how much information* to be processed are important for predicting instruction needs. This pattern is also found in the three best-performing experiments (Figure 4).

Exp. 3, was based on user-related characteristics only. Familiarity played the most important role here, followed by the sense of direction and egocentric preference scores. These observations are consistent with previous research on the importance of familiarity in activity recognition in wayfinding [39, 3] and the fact that instructions were given in an egocentric viewpoint. The same pattern for spatial strategy scores was also observed in **Exp. 3.3**. Among the personality traits, openness, which correlates with eagerness to learn and experience new things, is also the most important factor in **Exp. 3.2**. The relationship between openness and need for instruction is not well studied, but some studies have reported that individuals with higher openness tend to prefer more creative and less structured tasks [37] and therefore may need less detailed instructions. Our results suggest the same: Wayfinders with higher openness scores tend to show a lower need for instruction. However, further research is needed to decipher the relationship between this personality trait and prior experience (i.e., familiarity), task description and complexity, etc.

The result of **Exp. 4** and its sub-experiments with environmental features shows that segment-related features, including distance measurements to and from the immediate non-turn intersections and the previous and next turn points (which are the target of the instructions), as well as the time passed since onset, are the most important. This means that the longer the distance to these decision points the more probable it is to need instructions again. Analysis of this feature between the two familiarity conditions shows that distance to and from turn points is equally relevant for both conditions, but the distance to and from non-turn points is less dominant for familiar cases. These observations are consistent with those of [17], in which the authors showed that unfamiliar participants ask for instructions earlier on longer segments, and because the upcoming decision point is likely to be seen later on long segments, unfamiliar wayfinders might experience higher levels of uncertainty due to their less developed mental representation of the area. This explanation is also valid for our observation. All these measures that somehow capture the distance to/from different points on the route (time since start or length of segments) may also be related to the capacity of the wayfinder's working memory, which means that the longer the distance, the more likely it is that someone will forget the instruction and need it again. However, in our study, we did not control for this characteristic, and according to the psychological literature, working memory as a cognitive process of temporarily storing and processing information in the mind to perform tasks, while playing an important role in following instructions, is not the only factor that provides an advantage in a complex task environment [42]. Other factors that play a role in a complex task environment, such as problem-solving ability, creativity, or other

cognitive skills, may provide an advantage beyond good working memory [13]. However, we cannot relate our results to these explanations because our data do not provide corresponding information on these merits. Secondly important are PoI-based features, including the overall density of PoIs and *shop* in particular. Overall PoI density can be considered as a good indicator of the density of the urban landscape, which affects human-environment interactions including information seeking for wayfinding [31]. Our results suggest that the higher the PoI density, the higher the cognitive effort required to match the received instruction to the environment, and the greater the probability of requesting the same instruction again.

Finally, for the land use characteristics (e.g. *green urban area*, *roads*, etc.), which are equally important as PoIs in **Exp. 4** but not in **Exp. 5, 6, and 7**, no precise explanation can be offered because land use is likely biased by the study design. The familiar trials were conducted closer to the center of the city, while the unfamiliar trials were largely in the outskirts with different land use. This may justify the absence of land use characteristics in the experiments with combined features, as familiarity already encodes this difference. The effect of this environmental aspect needs to be further explored.

None of the experiments with a single category achieved practically high accuracy, with the exception of the environment category. Our combinatorial experiments aimed to find a pruned list of features that are most informative for predicting instruction demand. **Exp. 5, 6, and 7** are the best-performing experiments, and we consider the last experiment to be the optimal one with one-third of the features and only less than 1% loss in accuracy. In the last three experiments, we can see some similarities in the features. However, even in terms of features, we consider the permutatively selected features (listed in Figure 4 Exp. 7) to be the most informative features for predicting instruction needs. This list of features encodes well the wayfinding situation in terms of the instruction needs related to the wayfinder's relative position to the next and previous decision points (both turn and non-turn points); the effect of cognitive load on fixation behavior caused by two factors: the density of the urban landscape and PoIs in the environment and the processing of the instructional information (length and content); and finally, the characteristics of the user, from personality to preference for spatial strategies, gender, and familiarity.

7 Conclusion and Future Work

This paper presents the results of 15 ML experiments that predict the need for navigation instructions with an accuracy of 78.4%. The predictions are based on a combination of factors such as the wayfinder's position, the next decision points, cognitive load, the amount of visual information, the length and content of the instructions, and the user's personality traits and spatial strategies. The findings have theoretical and practical implications for better understanding the cognitive aspects of wayfinding and for adapting navigation instructions in real time.

The features used in the experiments encode several aspects of the wayfinder's situation: The wayfinder's position with respect to the starting point, the previous and upcoming decision points (both turn and non-turn), the cognitive load due to information processing and environmental perception reflected in fixation behavior and caused by the density of the landscape and the amount of visual information, the length and content of the given instruction, and finally the user's characteristics in terms of personality traits and spatial strategies. In our experimental design, instruction demand was defined as the frequency with which the same instruction is requested after it has been heard once. In other navigation systems with different HCI components, this demand can be defined, for example, as the

frequency of transition between the map screen and the environment, or as the number of fixations on the augmented information in AR-based systems after a first viewing. In any case, this prediction offers advantages from both theoretical and application perspectives: Behavior before and after an instruction is retrieved contains valuable information about the cognitive aspects of human-environment interaction and spatial perception. Knowing what external and internal features affect this demand can help us better understand the spatial-cognitive aspects of wayfinding and even mental states of uncertainty, being lost, or needing reassurance that are common in wayfinding but not yet well explored. A further research question would be when and in which stage of wayfinding we feel such needs more strongly and for which purpose (e.g. self-localization or route planning), the repeated instructions may be useful.

Since our prediction results are based on unseen data, it is very likely that a pre-trained model can perform on-the-fly predictions as a module of the navigation system, which can be beneficial for real-time instruction adaptation. Off-line predictions can also be used as a measurable metric for evaluating navigation instructions. However, further research is needed to examine the generalizability of our observations for different modalities.

References

- 1 Y. Abdelrahman, A. A. Khan, J. Newn, E. Velloso, Sh. Ashraf Safwat, J. Bailey, A. Bulling, F. Vetere, and A. Schmidt. Classifying attention types with thermal imaging and eye tracking. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 3(3), 2019. doi:10.1145/3351227.
- 2 N. Alinaghi and I. Giannopoulos. Consider the head movements! saccade computation in mobile eye-tracking. In *2022 Symposium on Eye Tracking Research and Applications*, 2022.
- 3 N. Alinaghi, M. Kattenbeck, and I. Giannopoulos. I can tell by your eyes! continuous gaze-based turn-activity prediction reveals spatial familiarity. In *15th Intl. Conf. on Spatial Information Theory (COSIT 2022)*. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2022.
- 4 N. Alinaghi, M. Kattenbeck, A. Golab, and I. Giannopoulos. Will you take this turn? gaze-based turning activity recognition during navigation. In *11th Intl. Conf. on Geographic Information Science (GIScience 2021)-Part II*. Leibniz-Zentrum für Informatik, 2021.
- 5 L. S. Ambati and O. El-Gayar. Human activity recognition: a comparison of machine learning approaches. *J. of the Midwest Association for Information Systems (JMWAIS)*, 2021(1):4, 2021.
- 6 T. Appel, N. Sevchenko, F. Wortha, K. Tsarava, K. Moeller, M. Ninaus, En. Kasneci, and P. Gerjets. Predicting cognitive load in an emergency simulation based on behavioral and physiological measures. In *2019 Intl. Conf. on Multimodal Interaction, ICMI '19*, pages 154–163, New York, NY, USA, 2019. Association for Computing Machinery.
- 7 A. D Baddeley, N. Thomson, and M. Buchanan. Word length and the structure of short-term memory. *J. of verbal learning and verbal behavior*, 14(6):575–589, 1975.
- 8 A. Brügger, K. Richter, and S. Fabrikant. How does navigation system behavior influence human behavior? *Cognitive research: principles and implications*, 4:1–22, 2019.
- 9 N. V Chawla, K. W Bowyer, L. O Hall, and W Ph. Kegelmeyer. Smote: synthetic minority over-sampling technique. *J. of artificial intelligence research*, 16:321–357, 2002.
- 10 L. De Cock, N. Van de Weghe, K. Ooms, I. Saenen, N. Van Kets, G. Van Wallendael, P. Lambert, and P. De Maeyer. Linking the cognitive load induced by route instruction types and building configuration during indoor route guidance, a usability study in vr. *Intl. J. of Geographical Information Science*, 36(10):1978–2008, 2022.
- 11 W. Dong, H. Liao, B. Liu, Z. Zhan, H. Liu, L. Meng, and Y. Liu. Comparing pedestrians' gaze behavior in desktop and in real environments. *Cartography and Geographic Information Science*, 47(5):432–451, 2020. doi:10.1080/15230406.2020.1762513.
- 12 M. Duckham, S. Winter, and M. Robinson. Including landmarks in routing instructions. *J. of location based services*, 4(1):28–52, 2010.


- 13 S. Dunham, E. Lee, and A. M Persky. The psychology of following instructions and its implications. *American J. of Pharmaceutical Education*, 84(8), 2020.
- 14 P. Fogliaroni, D. Bucher, N. Jankovic, and I. Giannopoulos. Intersections of our world. In *10th Intl. Conf. on geographic information science*, volume 114, page 3. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2018.
- 15 I. Giannopoulos, P. Kiefer, M. Raubal, K. Richter, and T. Thrash. Wayfinding Decision Situations: A Conceptual Model and Evaluation. In *Proc of GIScience 2014*, 2014.
- 16 F. Goebel, K. Kurzhals, V. R. Schinazi, P. Kiefer, and M. Raubal. Gaze-adaptive lenses for feature-rich information spaces. In *ACM Symposium on Eye Tracking Research and Applications*, ETRA '20 Full Papers, New York, NY, USA, 2020. Association for Computing Machinery. doi:10.1145/3379155.3391323.
- 17 A. Golab, M. Kattenbeck, G. Sarlas, and I. Giannopoulos. It's also about timing! when do pedestrians want to receive navigation instructions. *Spatial Cognition & Computation*, 22(1-2):74–106, 2022.
- 18 G. Gunzelmann, J. R Anderson, and S. Douglass. Orientation tasks with multiple views of space: Strategies and performance. *Spatial Cognition and Computation*, 4(3):207–253, 2004. doi:10.1207/s15427633scc0403_2.
- 19 H. He and E. A Garcia. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284, 2009.
- 20 H. Huang, M. Schmidt, and G. Gartner. Spatial knowledge acquisition with mobile maps, augmented reality and voice in the context of gps-based pedestrian navigation: Results from a field test. *Cartography and Geographic Information Science*, 39(2):107–116, 2012.
- 21 M. Adam Just and Patricia A. C. Eye fixations and cognitive processes. *Cognitive psychology*, 8(4):441–480, 1976.
- 22 M. Keskin and P. Kettunen. Potential of eye-tracking for interactive geovisual exploration aided by machine learning. *Intl. J. of Cartography*, pages 1–23, 2023. doi:10.1080/23729333.2022.2150379.
- 23 A. Klippel, H. Tappe, and Ch. Habel. Pictorial representations of routes: Chunking route segments during comprehension. In *Spatial Cognition III: Routes and Navigation, Human Memory and Learning, Spatial Representation and Spatial Learning 8*. Springer, 2003.
- 24 A. Klippel, H. Tappe, L. Kulik, and P. U Lee. Wayfinding choremes—a language for modeling conceptual route knowledge. *J. of Visual Languages & Computing*, 16(4):311–329, 2005.
- 25 A. Klippel and S. Winter. Structural salience of landmarks for route directions. In *Spatial Information Theory: Intl. Conf., COSIT 2005, Ellicottville, NY, USA, September 14-18, 2005. Proceedings 7*, pages 347–362. Springer, 2005.
- 26 J. Krukar, V. Joy Anacta, and A. Schwering. The effect of orientation instructions on the recall and reuse of route and survey elements in wayfinding descriptions. *J. of Environmental Psychology*, 68:101407, 2020.
- 27 A. Lakehal, S. Lepreux, L. Letalle, and Ch. Kolski. From wayfinding model to future context-based adaptation of hci in urban mobility for pedestrians with active navigation needs. *Intl. J. of Human–Computer Interaction*, 37(4):378–389, 2021.
- 28 H. Liao, W. Dong, H. Huang, G. Gartner, and H. Liu. Inferring user tasks in pedestrian navigation from eye movement data in real-world environments. *Intl. J. of Geographical Information Science*, 33(4):739–763, 2019.
- 29 B. Ludwig, G. Donabauer, D. Ramsauer, and K. al Subari. Urwalking: Indoor navigation for research and daily use. *KI - Künstliche Intelligenz*, 2023.
- 30 S. Münzer and Ch. Hölscher. Entwicklung und validierung eines fragebogens zu räumlichen strategien. *Diagnostica*, 2011.
- 31 C. Nothegger, S. Winter, and M. Raubal. Selection of salient features for route directions. *Spatial cognition and computation*, 4(2):113–136, 2004.

- 32 L. Pillette, G. Moreau, J. Normand, M. Perrier, A. Lecuyer, and M. Cogne. A systematic review of navigation assistance systems for people with dementia. *IEEE Transactions on Visualization and Computer Graphics*, 2022.
- 33 B. Rammstedt, Ch. Kemper, M. Céline Klein, C. Beierlein, and A. Kovaleva. Eine kurze skala zur messung der fünf dimensionen der persönlichkeit: big-five-inventory-10 (bfi-10). *Methoden, Daten, Analysen (mda)*, 7(2):233–249, 2013.
- 34 M. Raubal and S. Winter. Enriching wayfinding instructions with local landmarks. In *Intl. Conf. on geographic information science*, pages 243–259. Springer, 2002.
- 35 K. Richter, M. Tomko, and S. Winter. A dialog-driven process of generating route directions. *Computers, Environment and Urban Systems*, 32(3):233–245, 2008.
- 36 D. D. Salvucci and J. H. Goldberg. Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the 2000 Symposium on Eye Tracking Research & Applications*, ETRA '00, pages 71–78, New York, NY, USA, 2000. ACM. doi:10.1145/355017.355028.
- 37 D. Twomey, E. Burns, and Sh. Morris. Personality, creativity, and aesthetic preference: Comparing psychoticism, sensation seeking, schizotypy, and openness to experience. *Empirical Studies of the Arts*, 16(2):153–178, 1998.
- 38 P Unema. Differences in eye movements and mental work-load between experienced and inexperienced motor vehicle drivers. *Visual search*, pages 193–202, 1990.
- 39 J. M Wiener, S. J Büchner, and C. Hölscher. Taxonomy of human wayfinding tasks: A knowledge-based approach. *Spatial Cognition & Computation*, 9(2):152–165, 2009.
- 40 J. M Wiener, Ch. Hölscher, S. Büchner, and L. Konieczny. Gaze behaviour during space perception and spatial decision making. *Psychological Research*, 76(6):713–729, 2012. doi:10.1007/s00426-011-0397-5.
- 41 S. Winter, M. Tomko, B. Elias, and M. Sester. Landmark hierarchies in context. *Environment and Planning B: Planning and Design*, 35(3):381–398, 2008.
- 42 T. Yang. *The role of working memory in following instructions*. PhD thesis, University of York, 2011.
- 43 W. Zhang, X. Zhao, and Z. Li. A comprehensive study of smartphone-based indoor activity recognition via xgboost. *IEEE Access*, 7:80027–80042, 2019.
- 44 B. Zhu, J. G Cruz-Garza, Q. Yang, M. Shoaran, and S. Kalantari. Identifying uncertainty states during wayfinding in indoor environments: An eeg classification study. *Advanced Engineering Informatics*, 54:101718, 2022.

Transitions in Dynamic Point Labeling

Thomas Depian ✉

Algorithms and Complexity Group, TU Wien, Austria

Guangping Li ✉ 

Algorithm Engineering Group, TU Dortmund, Germany

Martin Nöllenburg ✉ 

Algorithms and Complexity Group, TU Wien, Austria

Jules Wolms ✉ 

Algorithms and Complexity Group, TU Wien, Austria

Abstract

The labeling of point features on a map is a well-studied topic. In a static setting, the goal is to find a non-overlapping label placement for (a subset of) point features. In a dynamic setting, the set of point features and their corresponding labels change, and the labeling has to adapt to such changes. To aid the user in tracking these changes, we can use morphs, here called *transitions*, to indicate how a labeling changes. Such transitions have not gained much attention yet, and we investigate different types of transitions for labelings of points, most notably *consecutive* transitions and *simultaneous* transitions. We give (tight) bounds on the number of overlaps that can occur during these transitions. When each label has a (non-negative) weight associated to it, and each overlap imposes a penalty proportional to the weight of the overlapping labels, we show that it is NP-complete to decide whether the penalty during a simultaneous transition has weight at most k . Finally, in a case study, we consider geotagged Twitter data on a map, by labeling points with rectangular labels showing tweets. We developed a prototype implementation to evaluate different transition styles in practice, measuring both number of overlaps and transition duration.

2012 ACM Subject Classification Theory of computation → Computational geometry; Human-centered computing → Geographic visualization

Keywords and phrases Dynamic labels, Label overlaps, Morphs, NP-completeness, Case study

Digital Object Identifier 10.4230/LIPIcs.GIScience.2023.2

Related Version *Full Version*: <https://arxiv.org/abs/2202.11562>

Funding *Guangping Li*: Funded by the Austrian Science Fund (FWF) under grant P31119.

Jules Wolms: Funded partially by the Austrian Science Fund (FWF) under grant P31119 and partially by the Vienna Science and Technology Fund (WWTF) under grant ICT19-035.

1 Introduction

Maps are ubiquitous in the modern world: from geographic to political maps, and from detailed road networks to schematized metro maps, maps are used on a daily basis. Advances in technology allow us to use digital maps on-the-fly and in a highly interactive fashion, by means of panning, zooming, and searching for map features. Besides changes induced by the user, maps can also change passively, for example automated panning during gps routing, or changing points of interest when visualizing time-varying geospatial (point) data.

Important features on a map are often labeled. Examples of such features are areas (such as countries and mountain ranges), curves (for example roads and rivers), and most importantly points (of interest). The aforementioned interactions force map features and their corresponding labels to change, by appearing, disappearing, or changing position. Instead of swapping between the map before and after such changes, we can use morphs, here called *transitions*, to allow the user to more easily follow changes in map features and labelings.



© Thomas Depian, Guangping Li, Martin Nöllenburg, and Jules Wolms;
licensed under Creative Commons License CC-BY 4.0

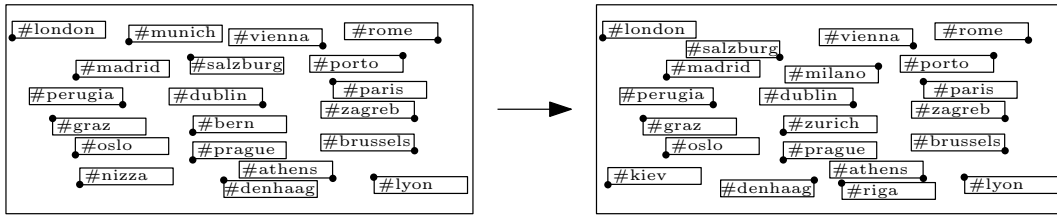
12th International Conference on Geographic Information Science (GIScience 2023).

Editors: Roger Beecham, Jed A. Long, Dianna Smith, Qunshan Zhao, and Sarah Wise; Article No. 2; pp. 2:1–2:19

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany



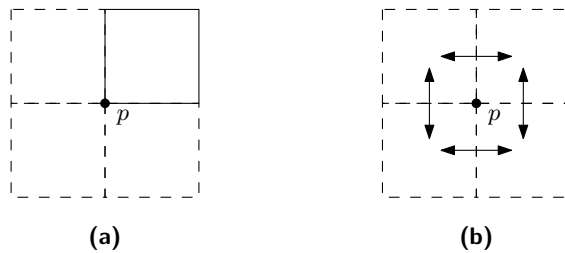
■ **Figure 1** A full visual scan of the individual labels is necessary to identify all changes [20].

Figure 1 shows why such transitions are important: even for two very similar point labelings, a lot of mental effort can be required to identify the differences.

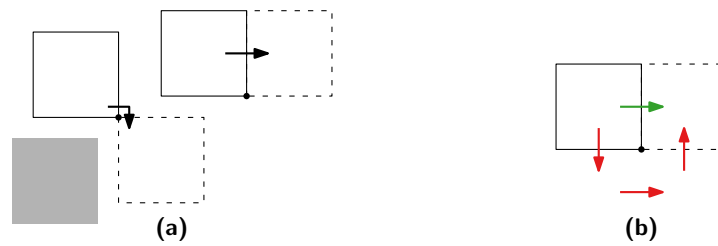
Automated map labeling is a well-researched topic within the geographic information science (GIS) and computational geometry community. In recent years, the GIS community has investigated the labeling of road networks [18], island groups [21], time-varying maps [2, 14], combining labeling with word clouds [5], and using human-in-the-loop approaches for labeling [13]. Algorithms have mainly focused on (the complexity of) computing labelings, in various static [1, 11, 22], interactive [3, 4, 12, 15], and dynamic or kinetic [6, 7, 9] settings.

In this paper we study transitions on maps that show point features P and their labels L . Let P be a finite point set in \mathbb{R}^2 , where each point $p_i \in P$ has a label $l_i \in L$ associated to it. Labels are axis-aligned rectangles in the frequently used four-position model, that is, each point p_i has four possible candidate positions to place label l_i [11] (see Figure 2a). While labels are often modeled as arbitrary (axis-aligned) rectangles, we use squares with side length $\sigma = 1$ for simplicity. In Appendix A we show how our results extend to arbitrary rectangles. A labeling $\mathcal{L} \subseteq L$ of P consists of a set of pairwise non-overlapping labels, and can be drawn on a map conflict-free, by drawing only the labels in \mathcal{L} with their associated points. If the label $l \in L$ for a point $p \in P$ is not contained in \mathcal{L} , we do not draw p either.

Furthermore, we work in a dynamic setting, where points appear and disappear at different moments in time, and the set P changes only through additions and deletions: the data we consider later consists of geotagged tweets, for which we know only the location at the moment they are tweeted, and hence data points do not move. Every time changes are made to P , a new overlap-free labeling must be computed, thus resulting in a change from labeling \mathcal{L}_1 , before the changes, to labeling \mathcal{L}_2 , afterwards. In this paper we study different types of transitions from \mathcal{L}_1 to \mathcal{L}_2 . During such a transition, the individual labels are allowed to move in the sliding-position model [22] (see Figure 2b). Our aim is to find transitions that achieve optimization criteria, such as minimizing the number of overlaps during a transition, or minimizing the time required to perform a transition. To our knowledge, this is the first time transitions have been studied in this way.



■ **Figure 2** (a) The four candidate positions for label l of point p , with l placed in the top-right position. (b) Labels continuously move between candidate positions using the sliding-position model.



■ **Figure 3** (a) Minimizing overlaps by moving around the gray stationary label. (b) Minimizing duration by using a single movement along the green arrow, instead of moving along the red arrows.

Problem description. Given two (overlap-free) labelings \mathcal{L}_1 and \mathcal{L}_2 , we denote a transition between them with $\mathcal{L}_1 \rightarrow \mathcal{L}_2$. Such a transition consists of changes of the following types.

Additions If only label l_i of a feature point p_i must be added, we denote this by $\mathcal{L}_1 \xrightarrow{A_i} \mathcal{L}_2$.

Removals If only label l_i of a feature point p_i must be removed, we denote this by $\mathcal{L}_1 \xrightarrow{R_i} \mathcal{L}_2$.

Movements If only label l_i of a feature point p_i must change from its position in \mathcal{L}_1 to a new position in \mathcal{L}_2 , we denote this by $\mathcal{L}_1 \xrightarrow{M_i} \mathcal{L}_2$. Movements are unit speed and axis-aligned, in the sliding-position model. Note that a diagonal movement, as in Figure 3a (left), is composed of a horizontal and a vertical movement, and hence takes two units of time.

A label is *stationary* if it remains unchanged during a transition. Applying multiple transitions consecutively is indicated by chaining the corresponding transition symbols: $\mathcal{L}_1 \xrightarrow{M_i M_j} \mathcal{L}_2$ denotes that label l_i moves before label l_j . Furthermore, $\mathcal{L}_1 \xrightarrow{M} \mathcal{L}_2$ is a shorthand for applying all movement-transitions simultaneously. All these notions extend to additions and removals, using A and R , respectively, instead of M . A transition has no effect if no point must be transformed with the respective transition, e.g., even if there are no additions, the transition $\mathcal{L}_1 \xrightarrow{A} \mathcal{L}_2$ is still applicable; it simply does not modify the labeling.

We aim to identify types of transitions that try to achieve the following goals.

\mathcal{G}_1 – Minimize overlaps While the two labelings are overlap-free, overlaps can occur during the transition from \mathcal{L}_1 to \mathcal{L}_2 . When too many overlaps happen at the same time, certain labels may (almost) completely disappear behind others during a transition, which defeats the purpose of the transition: allowing users to follow changes in the labeling. Thus, those overlaps should be avoided as much as possible, by, for instance, adjusting the movement direction of labels, as shown in Figure 3a.

\mathcal{G}_2 – Minimize transition duration Our main goal is to show a map in a (mostly) static state. However, we do not want to instantly swap \mathcal{L}_1 for \mathcal{L}_2 , since the user will have difficulties tracking all changes [20]. Though, a transition that takes too long can also cause users to lose attention [19]. Hence, we want transitions that can be completed in a short amount of time. This can be achieved by disallowing detours, as in Figure 3b, or by performing the changes simultaneously.

Note that ideally, one would also try to minimize the number of moving labels, as studies have showed that the amount of information humans can process is limited [17]. However, in this paper, we assume that we are given the new labeling \mathcal{L}_2 , which thus dictates the labels that have to move. Therefore, we see the task of computing a *stable* labeling \mathcal{L}_2 , i.e., one where only a few labels move, as an interesting research question in its own right.

Optimizing both goals \mathcal{G}_1 and \mathcal{G}_2 simultaneously is often impossible as there can be a trade-off: performing the transition as fast as possible to achieve \mathcal{G}_2 often leads to unnecessary overlaps, while preventing as many overlaps as possible to achieve \mathcal{G}_1 may require more time.

However, to work towards both \mathcal{G}_1 and \mathcal{G}_2 , we can perform all additions simultaneously, as well as all removals. Furthermore, if we perform removals before movements, and movements before the additions, we create free space for the movements, to reduce the number of overlaps without wasting time. Let X be an arbitrary way of performing all movements required to change from \mathcal{L}_1 to \mathcal{L}_2 (consecutively or simultaneously), then we can observe the following.

► **Observation 1.1.** *A transition of the form $\mathcal{L}_1 \xrightarrow{RXA} \mathcal{L}_2$ aids in achieving both \mathcal{G}_1 and \mathcal{G}_2 .*

We introduce two overarching *transition styles* in this paper: *consecutive transitions* and *simultaneous transitions*. Each such transition style is a variant of the style RXA , as prescribed by Observation 1.1, and fills in the movement described by X in a unique way. For a consecutive transition the movement X consists of a sequence of label movements, whereas for a simultaneous transition we have $X = M$. These transition styles each incur different transition durations. Since we expect a trade-off between \mathcal{G}_1 and \mathcal{G}_2 , we specifically analyze the number of overlaps during transitions of the two styles.

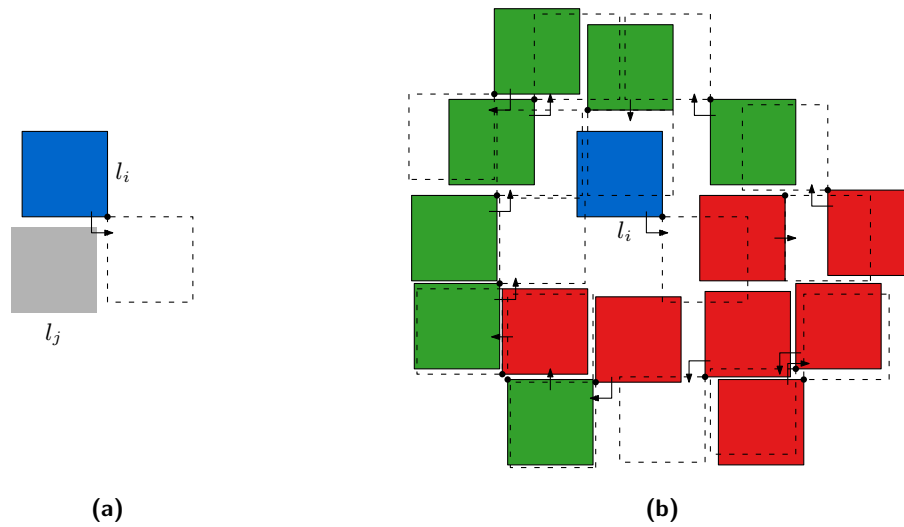
Related work. Our problem description resembles earlier work on point labeling, but it also has subtle differences. For example, the optimization criteria we care for, minimizing overlaps and time required for labels to move, were already investigated by de Berg and Gerrits [9]. They showed that there often is a clear trade-off between these criteria when dealing with moving labels. However, in their model, points are allowed to move (even during label movement), while our points are static, and change only through additions and deletions. Furthermore, in the PSPACE-hardness framework by Buchin and Gerrits [7] points are often static and only labels move. Hence, their dynamic labeling instances are similar to transitions. Though, a distinct difference is that labels must be allowed to move back and forth in the dynamic labeling instances of the hardness reduction. Since we disallow detours in transitions (see goal \mathcal{G}_2), this reduction is not easily transferred to our setting.

Finally, our analysis of the number of overlaps in transitions draws multiple parallels with the analysis of topological stability, introduced in the framework for algorithm stability [16]. This framework provides various (mathematical) definitions of stability for algorithms on time-varying data: Intuitively, small changes in the input of an algorithm should lead to small changes in the output. Topological stability prescribes that the output changes continuously. The (topological) stability ratio of an algorithm then measures how close to optimum the stable output is: when an optimal solution undergoes a discrete change, a topologically stable output has to continuously morph through suboptimal solutions. Similarly, we analyze transitions with continuous movement of various styles. We then analyze how close to overlap-free a labeling is during a transition by counting overlaps.

Contributions. In Sections 2 and 3 we analyze the worst-case number of overlaps of consecutive and simultaneous transitions, respectively. In Section 3 we additionally consider instances where we associate weights to the labels (and to their overlaps) and prove that it is NP-hard to minimize the weight of overlaps in simultaneous transitions. Finally, in Section 4 we investigate in a case study how the transition styles perform on the described goals.

2 Consecutive Transitions

Naive transitions. Before we can propose more elaborate transition styles, we first evaluate the potential overlaps for a single label performing its movement. Figure 4a shows how only a single stationary square label can interfere with the moving label.



■ **Figure 4** (a) Since all labels are squares with side length σ , the moving blue label l_i can overlap only a single gray stationary label l_j . (b) The blue label l_i overlaps 14 other labels during the movement transitions. The green labels move before l_i , red labels move after l_i .

► **Lemma 2.1.** *In $\mathcal{L}_1 \xrightarrow{RM_iA} \mathcal{L}_2$, where only label l_i moves, at most one overlap can occur.*

Proof. As we perform removals before the movement and additions afterwards, we can guarantee that the start and end positions of label l_i are free. Thus any overlap can occur only during diagonal movement of l_i , when l_i moves from one candidate position in \mathcal{L}_1 , to a non-adjacent candidate position in \mathcal{L}_2 . Assume without loss of generality that l_i traverses the lower-left label position, when moving from top-left to bottom-right. Only a single other (stationary) label l_j can be positioned such that both \mathcal{L}_1 and \mathcal{L}_2 are overlap-free and the label overlaps with the area traversed by l_i (see Figure 4a). Any additional label overlapping the traversed area, without overlapping l_j , would overlap the start or end position of l_i . ◀

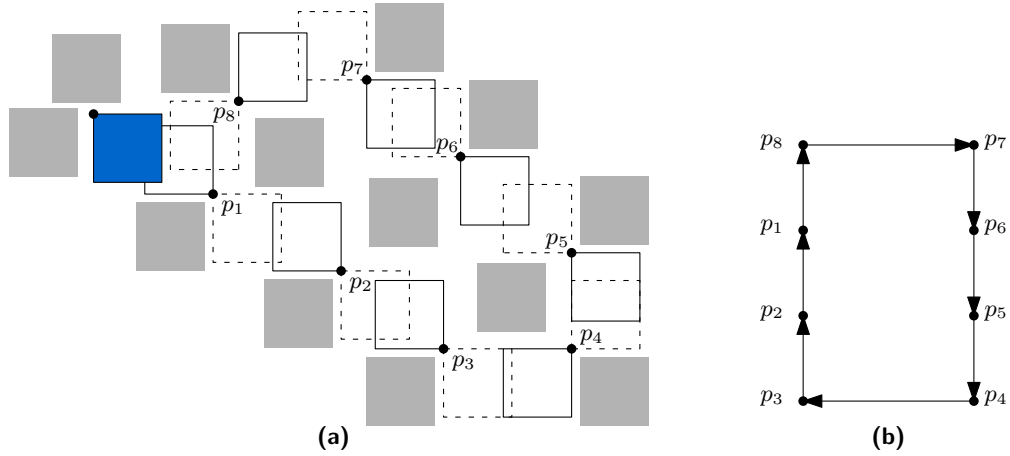
Next we consider an arbitrary order of all n moving labels in a transition. We define a conflict graph, which has a vertex for each moving label, and an edge between overlapping labels. With a packing argument we locally bound the degree of each of the n moving labels to 14 by considering the start, intermediate, and end position of such a label (these overlaps are achieved in Figure 4b). By the handshaking lemma this results in at most $7n$ overlaps. For more details on the proof of Lemma 2.2 see the full version [10].

► **Lemma 2.2.** *In $\mathcal{L}_1 \xrightarrow{RM_1 \dots M_n A} \mathcal{L}_2$ at most $7n$ overlaps can occur.*

DAG-based transitions. To refine the naive approach, we model dependencies between movements in a *movement graph*, and use it to order movements and avoid certain overlaps.

► **Definition 2.3** (Movement graph). *Let $\mathcal{M} = \{M_1, \dots, M_n\}$ be a set of movements. Create for each movement $M_i \in \mathcal{M}$ a vertex v_i , and create a directed edge from v_i to v_j , $v_i \rightarrow v_j$, if some intermediate or end position of M_j overlaps with the start position of M_i , or the end position of M_j overlaps with some intermediate position of M_i : In both cases movement M_i should take place before movement M_j . If intermediate positions of M_i and M_j overlap, create the edge $v_i \rightarrow v_j$, $i < j$. This results in the movement graph $G_{\mathcal{M}}$ (see Figure 5).*

A *feedback arc set* in a movement graph is a subset of edges that, when removed, breaks all cycles, resulting in a directed acyclic graph (DAG). We order movements using this DAG.



■ **Figure 5** (a) The blue label is added in this transition and forces $n + m$ inevitable overlaps during movement ($n = 8$ and $m = 1$). Gray labels are stationary. (b) The corresponding movement graph.

► **Theorem 2.4.** *Movements in $\mathcal{L}_1 \xrightarrow{RM_1 \dots M_n A} \mathcal{L}_2$ can be rearranged such that at most $n + m$ overlaps occur, if $G_{\mathcal{M}}$, with $\mathcal{M} = \{M_1, \dots, M_n\}$, has a feedback arc set of size m .*

Proof. By Lemma 2.1, we know that at most one overlap occurs when moving a single label to a free end position. This leads to at most n overlaps for n consecutively moving labels, if no label moves to (or through) a position occupied by a label, which starts moving later.

Let $G_{\mathcal{M}}$ be a movement graph with $\mathcal{M} = \{M_1, \dots, M_n\}$. There are two cases:

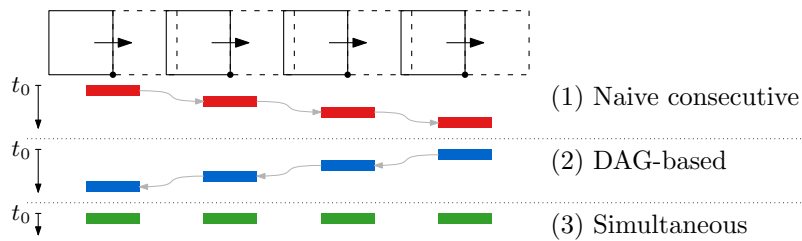
Case (1) If $G_{\mathcal{M}}$ is acyclic, then handling all movements according to any topological ordering of the vertices of $G_{\mathcal{M}}$ produces no additional overlaps.

Case (2) If $G_{\mathcal{M}}$ contains cycles, then overlaps may be inevitable because each label in such a cycle wants to move to or through a position that is occupied by another moving label. Moreover, as the movements happen consecutively, one label in this cycle must move first and therefore may cause an overlap. Let m be the smallest number of edges that must be removed to break each cycle in $G_{\mathcal{M}}$, i.e., the size of a minimum feedback arc set S . As $G_{\mathcal{M}}$ is cycle-free after removing S , case (1) applies and m additional overlaps suffice. ◀

We can see in Figure 5 that this bound is tight. Furthermore, it is not always necessary to perform all movements consecutively. We can observe that movements which are unrelated in $G_{\mathcal{M}}$ can be performed simultaneously: when no overlap is possible, there is no edge in $G_{\mathcal{M}}$.

3 Simultaneous Transitions

Figure 6 shows three timelines of different transition styles, (1) a naive consecutive transition, (2) a DAG-based transition, and (3) a simultaneous transition. All transition styles start at some time t_0 and the order of the movements of the labels for (1) and (2) is indicated with gray arrows. While (1) produces four overlaps and takes four units of time, (2) and (3) produce no overlaps, and (3) only takes a single unit of time. This shows that it is sometimes unnecessary to perform the movements consecutively to minimize overlaps. In this section, we investigate both how simultaneous movements influence the number of overlaps, and the complexity of minimizing overlaps.



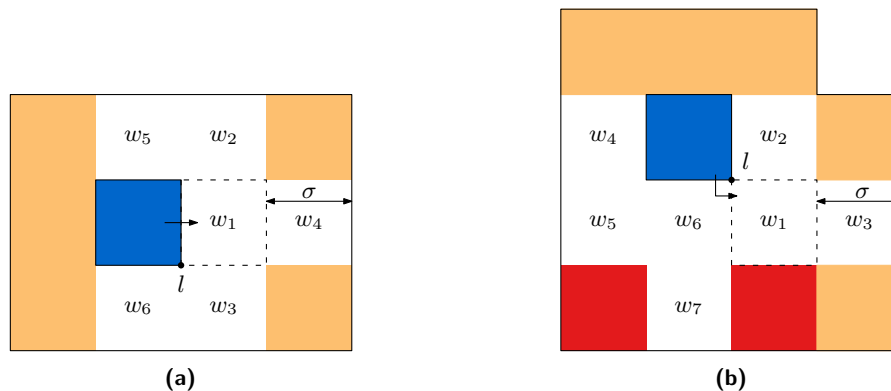
■ **Figure 6** Comparison of possible movement orderings with respect to \mathcal{G}_1 and \mathcal{G}_2 .

► **Theorem 3.1.** In $\mathcal{L}_1 \xrightarrow{RMA} \mathcal{L}_2$ at most $6n$ overlaps can occur, where n is the number of labels that must be moved, and all movements are performed at unit speed.

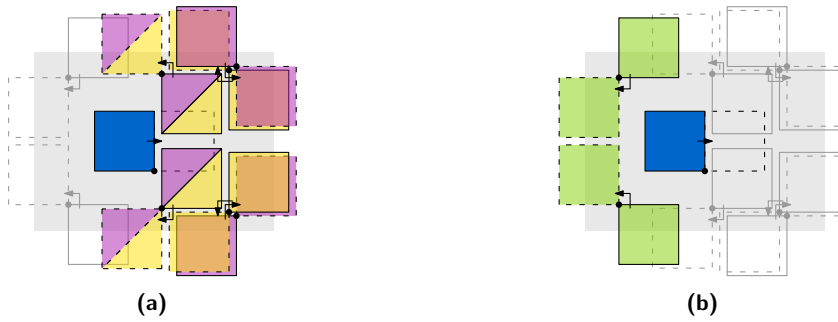
Proof. Let $\sigma = 1$ denote the side length of a label. To show that the total number of overlaps is at most $6n$, we model the overlaps in a graph and consider the neighborhood of individual vertices. Let G be a conflict-graph where each vertex v_i corresponds to a label l_i . If two labels l_i and l_j overlap during the transition, we create an edge (v_i, v_j) , i.e., each edge corresponds to an overlap. Observe that each edge is adjacent to at least one moving label since two stationary labels cannot overlap. We proceed by evaluating in G the maximum possible degree of a moving label l and restrict ourselves to a σ -wide border around the bounding box of the movement area of l , that represents the area other labels must touch (before the transition) to overlap with l . We call this area the *overlapping region* of l and it is illustrated in Figure 7. Labels not intersecting the overlapping region of l by construction cannot overlap with l . We proceed by considering the two possible types of movements for l .

Non-diagonal movement of l . For a label l that performs a non-diagonal movement, the overlapping region is illustrated in Figure 7a. The light-orange area in the overlapping region indicates that the start position of a label overlapping l cannot lie solely in this area. If a label starts in the area behind l , then such a label would never overlap with l , since labels move simultaneously. The start position of labels overlapping l can neither overlap only the $\sigma \times \sigma$ tiles diagonally adjacent to the end position of l , as the end position of those labels would overlap the end position of l , for any movement that allows the labels to overlap l .

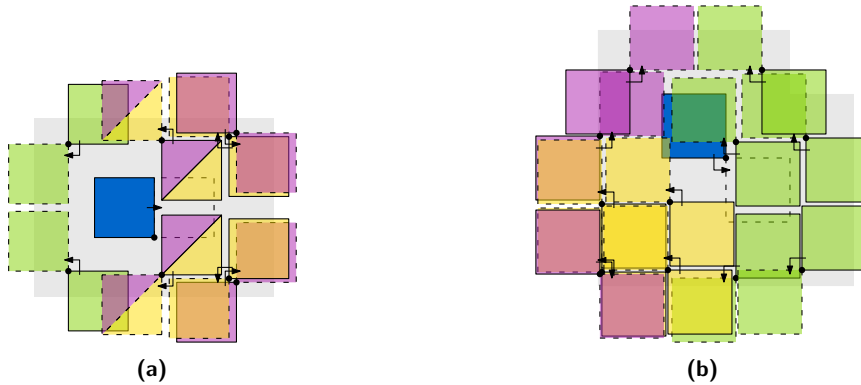
Next consider the remaining (white) area in the overlapping region, and see Figure 8 for our construction. Consider first the end position w_1 of l and the three $\sigma \times \sigma$ tiles w_2, w_3 and w_4 adjacent to it. The total height of w_1, w_2 and w_3 combined is 3σ , and hence the



■ **Figure 7** Overlapping regions for (a) non-diagonal and (b) diagonal movement of the blue label l .



■ **Figure 8** Labels overlapping with the blue label l that are located on the white tiles (a) $w_1 - w_4$ and (b) $w_5 - w_6$. The overlapping region of l is indicated in gray.



■ **Figure 9** Labels overlapping the blue label l that performs a (a) non-diagonal and a (b) diagonal movement. The overlapping region of l is indicated in gray.

start positions of at most four labels can be stacked vertically to overlap this area (see the labels with the color \blacksquare in Figure 8a). Similarly, w_1 and w_4 have a combined width of 2σ and height σ . Since the end position of l is adjacent to the start position of l we can put at most two labels horizontally next to each other in this area, while keeping \mathcal{L}_1 overlap free. However, as the height is σ we can stack at most two layers of such labels vertically (the \blacksquare labels in Figure 8a). As a result, w_1, w_2, w_3 and w_4 can together overlap with at most six start positions of other labels. Each label results in at most one overlap, and there is a movement direction for each label that achieves such an overlap, as shown in Figure 8a.

Now consider the $\sigma \times \sigma$ tiles w_5 and w_6 above and below the start position of l , respectively. We can place two labels, the ones colored \blacksquare in Figure 8b, such that their start positions overlap either of w_5 and w_6 . For example, for w_5 such labels can move diagonally down-left, to overlap l . In this case, it is impossible for a label overlapping w_6 to both overlap l and have an overlap-free end position. Conversely, we can place one label on w_5 and w_6 each and allow them both to move towards l , while ensuring overlap-free end positions (see Figure 8b). Observe that it is impossible to place two labels on both w_5 and w_6 in the latter case, as the vertical positioning that ensures overlap-free end positions of the labels, requires the labels to start farther from l . Those labels have to move a vertical distance of at least $\sigma/2$ to reach l , and hence also require a horizontal overlap of at least $\sigma/2$ with the start position of l , as l will have moved $\sigma/2$ rightwards before the other labels reach the start position of l . Thus, at most eight labels can overlap with l , and consequently the degree of the corresponding vertex is bounded by eight. See Figure 9a for the complete situation.

Diagonal movement of l . For diagonal movements, we consider w.l.o.g. the case where l performs a diagonal movement from top-left to bottom-right through the bottom-left corner. The overlapping region enlarges, as shown in Figure 7b. We can again eliminate the light-orange areas, as they mark areas that the start position of other labels cannot overlap exclusively, if they should overlap with l . As before, these areas are located behind the start position of l , and diagonally adjacent to the end position of l . The red areas are also eliminated, see the full version [10] for more details. We now repeat the process of filling the remaining (white) $\sigma \times \sigma$ tiles, w_1 to w_7 , with start positions for labels that can overlap with l during movement. The analysis is very similar to the non-diagonal case, with one exception: l can overlap with one stationary label, which can now occupy w_6 . We again do a case distinction on the possible label placements, showing that at most nine moving labels and one stationary label can overlap with l , or at most 12 moving labels can overlap with l . The upper bound of 12 overlaps for one label is tight (as shown in Figure 9b). See the full version [10] for the remaining details of this case.

Deriving an upper bound. To find an upper bound on the number of overlaps, consider the subgraph $G[V_M]$ induced by the set V_M of vertices that represent moving labels. The degree of one such vertex in $G[V_M]$, which represents a label l , is bounded by nine, in case l moves diagonally and a stationary label is present on the intermediate position of l , or by 12, otherwise. Both these bounds are higher than in the non-diagonal movement case, which would result in a degree of at most eight. Hence, in the worst case we have a degree sum between $9n$ and $12n$, respectively, since $|V_M| = n$. By the handshaking lemma, we then have between at most $\lceil \frac{9}{2} \rceil n$ and $6n$ edges in $G[V_M]$, respectively.

If we consider the original graph G , we can observe that it differs from $G[V_M]$, in terms of edges, only by the edges that are incident to a moving label and a stationary label. This means that the former case may result in more overlaps: As we have seen in one case of the above proof, and due to Lemma 2.1, a moving label can overlap with at most one stationary label. Since each of the edges in $E(G) \setminus E(G[V_M])$ is incident to exactly one vertex that represents a moving label, and we have n of such labels, $|E(G) \setminus E(G[V_M])|$ is bounded by n and consequently, we have at most $(\lceil \frac{9}{2} \rceil + 1)n$ overlaps with the stationary label present. However, in the worst case we still have an upper bound of at most $6n$ overlaps. ◀

3.1 Complexity of Computing Simultaneous Transitions

In this section, we show that it is NP-complete to minimize the number of overlaps in a weighted $\mathcal{L}_1 \xrightarrow{RMA} \mathcal{L}_2$ -transition by choosing the direction of diagonal movements.

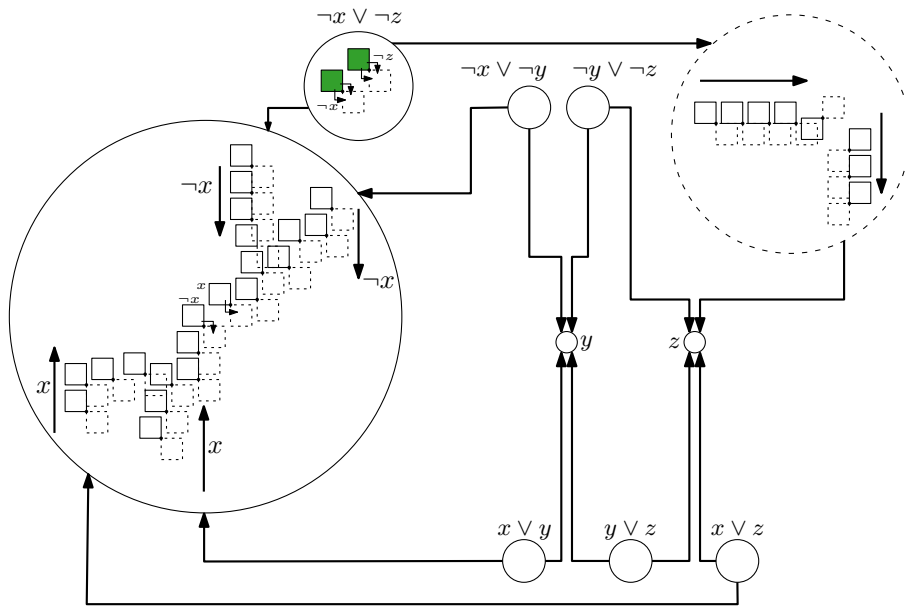
► **Definition 3.2** (Weighted Transition). *Let $\mathcal{L}_1 \xrightarrow{\Sigma} \mathcal{L}_2$ be a transition, where Σ denotes an arbitrary transition style of additions, movements, and removals, and let w be a weight function that assigns to each label $l \in L$ a non-negative weight $w(l) \in \mathbb{R}_0^+$. A weighted transition $\mathcal{L}_1 \xrightarrow[w]{\Sigma} \mathcal{L}_2$ performs $\mathcal{L}_1 \xrightarrow{\Sigma} \mathcal{L}_2$, but when two labels l_i and l_j overlap, a penalty of weight $w(l_i) \cdot w(l_j)$ is introduced. The total penalty W is equal to the sum of penalty weights.*

► **Problem 3.3.** *Given a weighted transition $\mathcal{L}_1 \xrightarrow[w]{RMA} \mathcal{L}_2$ and $k \in \mathbb{R}_0^+$, can we assign a movement direction to each diagonal movement such that the total penalty W is at most k ?*

We sketch the proof of Theorem 3.4 here, details can be found in the full version [10].

► **Theorem 3.4.** *It is NP-complete to decide whether W is at most k for $\mathcal{L}_1 \xrightarrow[w]{RMA} \mathcal{L}_2$.*

Proof sketch. Given a movement direction for each label, it is easy to check whether W is at most k by considering each pair of labels and checking for overlaps. Hence Problem 3.3 is contained in NP. For NP-hardness, we reduce from an instance F of PLANAR MONOTONE MAX 2-SAT [8]. Figure 10 gives an overview of the required gadgets. Clause and variable gadgets consist of two opposing labels at their core, corresponding, respectively, to the assignments of the two literals in a clause, or the binary choice for a variable. For an unsatisfied clause, an overlap occurs inside the clause gadget, whenever both labels move towards each other (inwards). The corresponding labels have weight one, and hence such an overlap would incur a penalty of weight one. A variable gadget has two opposing labels for setting the variable to *true* or *false*. Choosing a movement direction outward from the variable gadget, for example on the “true”-side, will cause a domino effect, propagating towards the gadgets of clauses with negative occurrences of this variable. There it results in inward movement, and hence this corresponds to setting the variable to not be false (and thus be true). Choosing the outward movement for both variable states is never beneficial: that variable is neither true nor false. The movement directions chosen in the variable gadgets are propagated to the appropriate clauses using the (planar) embedding of the incidence graph of F . All labels outside of clause gadgets have weight $n + 1$ and hence producing an overlap outside of a clause gadget will result in a large penalty of weight greater than n . As such, we either have movement directions that produce a total penalty of at most k for some positive $k < n$, and overlaps correspond to unsatisfied clauses, or we have a total penalty of at least n , and no clauses can be satisfied (or the variable assignment is inconsistent). Thus, $n - k$ clauses are satisfiable in F , if and only if we have k overlaps in our reduced instance. ◀



■ **Figure 10** Reduced instance for the formula $F = (\neg x \vee \neg z) \wedge (\neg x \vee \neg y) \wedge (\neg y \vee \neg z) \wedge (x \vee y) \wedge (y \vee z) \wedge (x \vee z)$. The weight of white and green labels is $n + 1$ and 1, respectively.

Figure 10 shows an example of the complete setup. There we reduce the formula $F = (\neg x \vee \neg z) \wedge (\neg x \vee \neg y) \wedge (\neg y \vee \neg z) \wedge (x \vee y) \wedge (y \vee z) \wedge (x \vee z)$ onto our problem. Circles with solid borders indicate the individual parts of the formula, while the dashed circle shows part of a transportation gadget. The big empty circles represent the clauses and small empty

■ **Table 1** The Keywords and #Hashtags we used to query the tweets.

corona	#corona	covid	#covid	covid19	#covid19	covid-19
vaccine	#vaccine	quarantine	#quarantine	lockdown	#lockdown	moderna
outbreak	#outbreak	immune	#immune	immunity	#immunity	biontech
who	#who	desease	#desease	masks	#masks	pfizer
pandemic	#pandemic	mutation	#mutation	ffp2	#ffp2	
#StaySave	#StayAtHome	#FlattenTheCurve		astrazeneca	johnson & johnson	

circles represent the variables. The arrows outside circles represent the transportation-gadgets that connect the individual parts according to the direction of the arrows. As there exists no variable assignment which satisfies F , we cannot achieve $W = k = 0$ but must encounter at least one overlap (and hence $W \geq 1$). This overlap occurs, for example, in the clause gadget for $(\neg x \vee \neg z)$ to achieve $W = 1$. Note that if we would try to resolve this overlap by, for instance, setting x to true and false at the same time, an overlap with penalty $(n + 1)^2 = 49$ would occur, for example, in the variable-gadget for the variable x .

4 Case Study: Twitter Data

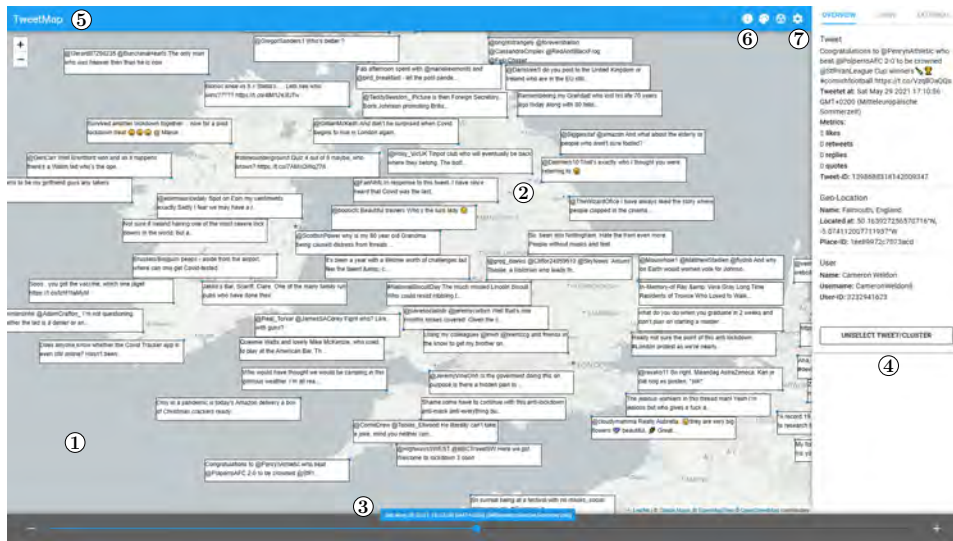
We implemented a prototype to analyze how the transition styles perform in a practical setting. In this prototype, we show geotagged tweets (Section 4.1) as rectangular labels on an interactive map (Section 4.3). In order to show an appropriate amount of information inside the labels, we use uniform-sized axis-aligned rectangles for the labels instead of squares. As all labels will have the same size, our theoretical results derived for square labels with side length 1 carry over to this model (by scaling horizontally). Finally, we measure the number of overlaps and the transition duration of the presented transition styles (Section 4.4).

4.1 Dataset

For our dataset, we queried 100,000 geotagged tweets related to the COVID-19 pandemic during the month of May 2021, see Table 1. After filtering and cleaning this dataset, 99,982 usable tweets were left. A tweet will be represented as a point $p \in P$, and as its label we display parts of the *Tweet Text* in a rectangular box, possibly truncating it if it is too long. For the spatial and temporal property we use the *Tweet Location* and *Tweet Date and Time* fields, respectively. The former field is a location attached to a tweet, which will determine the coordinates of the point $p \in P$. This is not necessarily a concrete location, but can be a (rectangular) area on the map. We choose an arbitrary location inside this area to prevent artificial cluster creation. The latter field defines the date and time the tweet was posted.

4.2 Dynamic Labeling Model

The dataset described above has spatiotemporal properties: each point $p \in P$ has a location and a time associated to it. Starting from this point in time, we consider p (and its associated tweet) *relevant* for three hours. The relevant tweets at a particular *time of interest* will form the set P of points that we want to label. Changes to the time of interest (dynamically) alter the set P of relevant tweets through additions and removals. Furthermore, in our implementation not all points in P will be in view at all times: For example, when the user zooms in on a particular part of the map, some points will be outside the view port. In such cases, we label only the subset $S \subseteq P$ of points that are inside the view port.



■ Figure 11 Screenshot of the prototype.

4.3 Implementation Details

The prototype computes a labeling in the four-position model of the relevant points P . Figure 11 shows a screenshot of our prototype. The main view area (1), in the center of the screen, shows a map and a labeling overlay. Furthermore, it contains blue dots (2) indicating the locations of the subset $S \subseteq P$. Below the map, the time slider (3) shows the currently selected time of interest. The side drawer on the right (4) shows further information of a tweet, if the user selects one. The top bar (5) allows the user to retrieve additional information about the map (6) and alter its state using the cogwheel (7).

The user can interact with the prototype by means of *panning* and *zooming* the map, as well as changing the time of interest by using the *time slider*. Panning is done by dragging the map using the mouse, while zooming is controlled using either the mouse wheel or the zoom indicators in the upper-left corner of the map. Zoom level changes step-wise. The time of interest is changed by dragging the indicator in the time slider, or using the + and - buttons on either side of the slider. Panning and zooming change the subset $S \subseteq P$ of points in view, while changes to the time of interest alter the relevant points P .

Computing a labeling. For a given subset $S \subseteq P$ of tweets that are both relevant and in view, we compute a labeling as follows. We create a conflict graph of labels for S and use a simple greedy approximation algorithm for a MAXIMAL INDEPENDENT SET I in this graph: iteratively add a minimum-degree vertex to I , that shares no edge with vertices in I .

When the user now interacts with the map, we again perform the same algorithm to find a new labeling, but we use two simple heuristics to improve the stability of the labeling. The first heuristic is based on desideratum D1 from [3]. Among other things, it proposes that the same labels should remain visible when zooming. To achieve this, we remove all neighbors of the previously shown labels in the conflict graph, to ensure they are picked again.

The second heuristic attempts to prevent unnecessary changes in the labeling: Let I_1 be the subset of labeled points that remained relevant and visible after panning/zooming/time change, and let I_2 be the newly computed set of points to be labeled. If I_2 is less than 2% larger than I_1 , then we simply keep the labeling of I_1 instead of swapping to a labeling of I_2 .

When the subset $S \subseteq P$ of relevant tweets in view is changed, through panning or zooming, or when P is dynamically altered by changes in the time of interest, our prototype will trigger a transition. Let $\mathcal{L}_1, \mathcal{L}_2$ be the computed labelings before and after the change, respectively. Our prototype supports naive, DAG-based, and simultaneous transitions for $\mathcal{L}_1 \rightarrow \mathcal{L}_2$. In each of these transition styles, a removal, an addition, or a movement of a single label has a duration of one second. A diagonal movement is split up into two non-diagonal movements with a duration of one second each, starting with the horizontal movement.

Naive transitions. Movements in this transition are performed consecutively in arbitrary order. Their order is based on the order in which we recognize the need for a movement.

DAG-based transitions. The movements in DAG-based transitions are also performed consecutively, though ordered according to a topological ordering of the movement graph $G_{\mathcal{M}}$. If $G_{\mathcal{M}}$ contains cycles, we remove the vertex with the lowest in-degree and first move the label of the removed vertex. Additionally, we perform unrelated movements in $G_{\mathcal{M}}$ simultaneously.

Simultaneous transitions. The movements in this transition are all performed simultaneously, immediately after the removals. The direction of diagonal movement is not optimized for minimum overlaps. Instead, we move horizontally first, to create a more uniform transition.

Implementation. The prototype is a three-tier-architecture, consisting of a graph-*database* (Neo4j) storing the tweets together with their potential label candidates, an *application tier* (Java Play Framework) computing the (new) labeling, and the *presentation tier* (Vue.js, Leaflet, and GreenSock) with which the user can interact and which visualizes the transitions.

In our case study we measured the running times of the individual components, which we report in Appendices B.1 and B.2. We can see that the majority of the time in the back-end (between 60% and 85%) is spent on querying the database, while the remaining parts run in less than 150ms in nearly all investigated cases. Computing the transitions in the front-end takes on average below 10ms, and never more than 20ms, which is negligible.

4.4 Measuring Transition Time and Number of Overlaps

In our case study, we use our prototype to simulate twelve interaction settings in six scenarios. The different scenarios we use in our case study are described in Table 2. In the first setting, we use different interactions depending on the scenario we consider. For each interaction type, we interact with the prototype by applying the following sequence of operations.

■ **Table 2** The different scenario states of the case study.

Scenario Name	Interaction	Map center		Zoom level	Time of interest
		Longitude	Latitude		
Italy	(a)	14.45	41.30	7	2021-05-29T13:20:00
Lausanne	(b)	6.37	46.45	7	2021-05-30T10:30:00
Leeds	(b)	-1.60	53.44	7	2021-05-29T13:00:00
Los Angeles	(c)	-117.78	33.84	9	2021-05-30T03:15:00
New Delhi	(a)	71.18	30.20	7	2021-05-29T08:30:00
Sao Paolo	(c)	-45.00	-20.65	7	2021-05-29T02:30:00

Interaction (a): (1) Increase the time of interest by 30 minutes, (2) zoom in by one zoom level with the help of the zooming indicators, (3) increase the latitude of the map’s center by 0.28 using the settings, and (4) increase the time of interest by five minutes with the + button next to the time slider.

Interaction (b): Interaction (a) in reverse order: (1) Increase the time of interest by five minutes with the + button next to the time slider, (2) increase the latitude of the map’s center by 0.28 using the settings, (3) zoom in by one zoom level with the help of the zooming indicators, and (4) increase the time of interest by 30 minutes

Interaction (c): (1) Zoom in by one zoom level with the help of the zooming indicators, *decrease* the time of interest by five minutes with the – button next to the time slider, (3) *decrease* the *longitude* of the map’s center by 1.7 using the settings, and (4) increase the time of interest by 20 minutes.

In Table 3, we report an overview of the most important results and refer to Appendix B for additional measurements. Simultaneous transitions are the fastest, and the naive transitions the slowest. Noteworthy is that the duration of the DAG-based transitions is close to the simultaneous transitions. This suggests that there is significant benefit in simultaneously performing movements that are unrelated in $G_{\mathcal{M}}$. Furthermore, we can see that the DAG-based transitions produce around half as many overlaps as the other styles, on average less than one overlap in each setting. Simultaneous transitions cause a similar number of overlaps as naive transitions, but on average the total number of overlaps is slightly better.

This case study shows that DAG-based transitions find a good compromise between the number of overlaps (\mathcal{G}_1) and the duration of the transitions (\mathcal{G}_2). However, simultaneous transitions are more appealing, if one favours faster transitions over the number of overlaps. Hence, we see a clear trade-off between the number of overlaps and transition duration, similar to previous work [9]. Videos of the different transition styles can be found online¹.

¹ Link to videos: https://osf.io/hnsvu/?view_only=7703ba40643440f8958a9b0120dc32f0

■ **Table 3** Evaluation results for each transition style in each setting, best scores per row are bold.

Scenario	Naive transitions			DAG-based transitions				Simultaneous transitions				
	#Overlaps			Duration [s]				#Overlaps				
	Avg.	Tot.	Max	Avg.	Avg.	Tot.	Max	Avg.	Avg.	Tot.	Max	Avg.
Italy, 1	0,40	2	6,50	2,20	0,20	1	4,49	1,40	0,60	3	2,50	1,00
Italy, 2	0,16	6	6,51	1,62	0,05	2	5,50	1,30	0,24	9	2,50	1,08
Lausanne, 1	0,40	2	16,49	6,50	0,40	2	4,51	2,30	0,40	2	2,51	1,50
Lausanne, 2	1,43	53	18,50	5,27	0,78	29	9,49	3,24	1,19	44	2,50	1,78
Leeds, 1	1,40	7	23,50	9,10	0,80	4	5,49	2,09	1,00	5	2,49	1,10
Leeds, 2	1,03	38	16,50	4,81	0,59	22	7,50	2,92	0,65	24	2,51	1,78
Los Angeles, 1	1,00	5	25,49	9,30	0,40	2	4,50	2,30	0,40	2	2,49	1,49
Los Angeles, 2	0,43	16	8,51	2,55	0,30	11	4,50	1,88	0,38	14	2,50	1,53
New Delhi, 1	1,20	6	22,48	10,80	0,60	3	8,49	4,60	1,40	7	2,50	2,00
New Delhi, 2	0,59	22	12,50	3,54	0,27	10	5,50	2,21	0,46	17	2,51	1,65
Sao Paolo, 1	0,20	1	25,50	9,70	0,00	0	8,49	3,30	0,60	3	2,50	1,50
Sao Paolo, 2	0,41	15	13,50	2,28	0,24	9	8,49	1,69	0,38	14	2,50	1,20
Avg. Setting 1	0.77	3.83	19.99	7.93	0.40	2.00	5.99	2.66	0.73	3.67	2.50	1.43
Avg. Setting 2	0.68	25.00	12.67	3.35	0.37	13.83	6.83	2.21	0.55	20.33	2.51	1.50

5 Conclusion

In this paper we performed a first investigation into the number of overlaps produced by transitions on labelings of points, and started by proving tight upper bounds for various transition styles. In addition, we implemented the transition styles in a prototype and performed a case study that revealed the need for sophisticated transition styles that find a good compromise between the number of overlaps and the duration of a transition. We see this paper as a first step towards understanding such transitions in point labeling. Therefore we have many open questions for future work, such as:

- Should we develop new transition styles or improve the existing ones? Can we utilize more structured movement, like performing all movements in the same direction simultaneously?
- Is it sensible to try to formalize more perception-oriented desiderata for transitions, such as the symmetry of transitions or the traceability of labels?
- Is choosing label directions in simultaneous transitions still NP-hard with unit weights?
- Can we compute a *stable* labeling \mathcal{L}_2 , that minimizes the number of moving labels?

References

- 1 Pankaj K. Agarwal, Marc J. van Kreveld, and Subhash Suri. Label placement by maximum independent set in rectangles. *Comput. Geom.*, 11(3-4):209–218, 1998. doi:10.1016/S0925-7721(98)00028-5.
- 2 Lukas Barth, Benjamin Niedermann, Martin Nöllenburg, and Darren Strash. Temporal map labeling: a new unified framework with experiments. In *Proc. 24th SIGSPATIAL*, pages 1–10, 2016. doi:10.1145/2996913.2996957.
- 3 Ken Been, Eli Daiches, and Chee-Keng Yap. Dynamic map labeling. *IEEE Trans. Vis. Comput. Graph.*, 12(5):773–780, 2006. doi:10.1109/TVCG.2006.136.
- 4 Ken Been, Martin Nöllenburg, Sheung-Hung Poon, and Alexander Wolff. Optimizing active ranges for consistent dynamic map labeling. *Comput. Geom.*, 43(3):312–328, 2010. doi:10.1016/j.comgeo.2009.03.006.
- 5 Sujoy Bhore, Robert Ganian, Guangping Li, Martin Nöllenburg, and Jules Wolms. Wordel: Aggregating point labels into word clouds. In *Proc. 29th SIGSPATIAL*, pages 256–267, 2021. doi:10.1145/3474717.3483959.
- 6 Sujoy Bhore, Guangping Li, and Martin Nöllenburg. An Algorithmic Study of Fully Dynamic Independent Sets for Map Labeling. In *Proc. 28th ESA*, pages 19:1–19:24, 2020. doi:10.4230/LIPIcs.ESA.2020.19.
- 7 Kevin Buchin and Dirk H. P. Gerrits. Dynamic Point Labeling is Strongly PSPACE-Complete. *Int. J. Comput. Geom. Appl.*, 24(4):373, 2014. doi:10.1142/S0218195914600127.
- 8 Kevin Buchin, Valentin Polishchuk, Leonid Sedov, and Roman Voronov. Geometric Secluded Paths and Planar Satisfiability. In *Proc. 36th SoCG*, pages 24:1–24:15, 2020.
- 9 Mark de Berg and Dirk H. P. Gerrits. Labeling Moving Points with a Trade-Off between Label Speed and Label Overlap. In *Proc. 21st ESA*, pages 373–384, 2013. doi:10.1007/978-3-642-40450-4_32.
- 10 Thomas Depian, Guangping Li, Martin Nöllenburg, and Jules Wolms. Transitions in Dynamic Map Labeling, 2022. doi:10.48550/arXiv.2202.11562.
- 11 Michael Formann and Frank Wagner. A packing problem with applications to lettering of maps. In *Proc. 7th SoCG*, pages 281–288, 1991.
- 12 Andreas Gemsa, Martin Nöllenburg, and Ignaz Rutter. Consistent Labeling of Rotating Maps. *J. Comput. Geom.*, 7(1):308–331, 2016. doi:10.20382/jocg.v7i1a15.
- 13 Fabian Klute, Guangping Li, Raphael Löffler, Martin Nöllenburg, and Manuela Schmidt. Exploring Semi-Automatic Map Labeling. In *Proc. 27th SIGSPATIAL*, pages 13–22, 2019. doi:10.1145/3347146.3359359.

- 14 Filip Krumpal. Labeling points of interest in dynamic maps using disk labels. In *Proc. 10th GIScience*, pages 8:1–8:14, 2018. doi:10.4230/LIPIcs.GISCIENCE.2018.8.
- 15 Chung-Shou Liao, Chih-Wei Liang, and Sheung H. Poon. Approximation algorithms on consistent dynamic map labeling. *Theor. Comput. Sci.*, 640:84–93, 2016. doi:10.1016/j.tcs.2016.06.006.
- 16 Wouter Meulemans, Bettina Speckmann, Kevin Verbeek, and Jules Wulms. A Framework for Algorithm Stability and Its Application to Kinetic Euclidean MSTs. In *Proc. 13th LATIN*, pages 805–819, 2018. doi:10.1007/978-3-319-77404-6_58.
- 17 George A. Miller. The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information. *Psychological review*, 63(2):81–97, 1956. doi:10.1037/h0043158.
- 18 Benjamin Niedermann and Martin Nöllenburg. An algorithmic framework for labeling road maps. In Jennifer A. Miller, David O’Sullivan, and Nancy Wiegand, editors, *Proc. 9th GIScience*, pages 308–322, 2016. doi:10.1007/978-3-319-45738-3_20.
- 19 Jakob Nielsen. *Usability Engineering*. Academic Press, 1993.
- 20 Ronald A. Rensink, John K. O’Regan, and James J. Clark. To See or not to See: The Need for Attention to Perceive Changes in Scenes. *Psychological Science*, 8(5):368–373, 1997. doi:10.1111/j.1467-9280.1997.tb00427.x.
- 21 Arthur van Goethem, Marc J. van Kreveld, and Bettina Speckmann. Circles in the water: Towards island group labeling. In *Proc. 9th GIScience*, pages 293–307, 2016. doi:10.1007/978-3-319-45738-3_19.
- 22 Marc J. van Kreveld, Tycho Strijk, and Alexander Wolff. Point labeling with sliding labels. *Comput. Geom.*, 13(1):21–47, 1999. doi:10.1016/S0925-7721(99)00005-X.

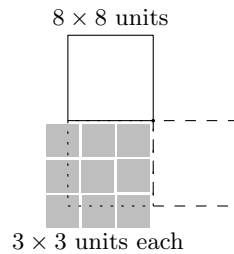
A Arbitrary Rectangle Labels

While in the main text we considered only square labels, point labelings often use arbitrary rectangles. If we allow our labels to be arbitrary rectangles, then it is no longer guaranteed that only one (stationary) label can overlap with the area traversed by the moving label. If we assume that the label with the largest side width (some $\sigma_{x_{\max}}$) must perform a diagonal movement, we can align $\frac{\sigma_{x_{\max}}}{\sigma_{x_{\min}}}$ stationary labels with a width of $\sigma_{x_{\min}}$, the smallest label width in our map, on the horizontal edge of the traversed area. As one label can always extend out of that traversed area without resulting in an invalid labeling \mathcal{L}_1 or \mathcal{L}_2 , we can put up to $\lceil \frac{\sigma_{x_{\max}}}{\sigma_{x_{\min}}} \rceil$ labels next to each other in the x -direction. The same holds for the y -axis, with maximum and minimum height $\sigma_{y_{\max}}$ and $\sigma_{y_{\min}}$, respectively. As we can put labels anywhere inside the traverse area, we can place up to $\lceil \frac{\sigma_{x_{\max}}}{\sigma_{x_{\min}}} \rceil \cdot \lceil \frac{\sigma_{y_{\max}}}{\sigma_{y_{\min}}} \rceil$ labels intersecting that area and therefore $\lceil \frac{\sigma_{x_{\max}}}{\sigma_{x_{\min}}} \rceil \cdot \lceil \frac{\sigma_{y_{\max}}}{\sigma_{y_{\min}}} \rceil$ overlaps occur during the movement (see Figure 12). This results in the following corollary, as an extension of Lemma 2.1.

► **Corollary A.1.** *When the labels are arbitrary rectangles with side length σ_{x_i} and σ_{y_i} , with $1 \leq i \leq n$ and n denotes the number of labels, performing transition $\mathcal{L}_1 \xrightarrow{RM_iA} \mathcal{L}_2$ for a label of a point p_i can result in at most $\lceil \frac{\sigma_{x_{\max}}}{\sigma_{x_{\min}}} \rceil \cdot \lceil \frac{\sigma_{y_{\max}}}{\sigma_{y_{\min}}} \rceil$ overlaps given that the end position of p_i is free, where $\sigma_{x_{\min}} = \min\{\sigma_{x_i} \mid 1 \leq i \leq n\}$ and $\sigma_{x_{\max}}$, $\sigma_{y_{\min}}$ and $\sigma_{y_{\max}}$ are defined similarly.*

Corollary A.1 shows how upper bounds on the number of overlaps produced by square labels can be extended to the setting of arbitrary rectangles. This introduces only a constant factor, depending on the ratio between the largest and smallest side lengths in each dimension. However, for many transitions adding the constant factor, as suggested by Corollary A.1, does not yield a tight bound. This stems from the fact that many upper bounds require overlaps with the start or end position of a label l , not just the traversed area of l . Since

those positions are solely occupied at respectively the beginning and the end of the transition, we cannot place $\lceil \frac{\sigma_{x_{\max}}}{\sigma_{x_{\min}}} \rceil \cdot \lceil \frac{\sigma_{y_{\max}}}{\sigma_{y_{\min}}} \rceil$ labels in those positions: many of those labels are unable to move away completely.



■ **Figure 12** The 8×8 label wants to perform a counterclockwise diagonal movement and overlaps with nine stationary 3×3 labels. The overlapping region is dotted, while the end position of the moving label is indicated with the dashed rectangle.

B Detailed Case Study Results

In this case study we measure both the running times of our implementation, as well as objective metrics (overlaps and transition duration) of the computed transitions. In the next two sections, we outline these two measurements separately.

For this case study we used a standard laptop with the following specifications and software versions.

- Intel®Core™ i5-8265U CPU @ 1.60GHz with 16 Gigabyte RAM
- Windows 11 Pro 21H2 (64 Bit) and Microsoft Edge Browser 109.0.1518.78
- Neo4j 4.1.7, Java openjdk 11.0.2, vue 3.0.11, leaflet 1.7.1, and gsap 3.6.1
- External 27" monitor with a resolution of 1920×1080px

B.1 Back-end computations

For the measurements of the back-end computations see Table 4.

B.2 Transition measurements

For the results of our measurements on the computed transitions see Table 5.

Table 4 Detailed results of the case study – Average running times in the back-end. QT = Query Time Database, CG CT = Conflict Graph Creation Time, MIS CT = MAXIMAL INDEPENDENT SET Computation Time.

Scenario	Setting	Conflict Graph (CG)				Changes in the Labeling				Running times [ms]			
		#Vertices	#Edges	#Added	#Removed	#Moved	QT	CG CT	MIS CT	Total			
Italy	1	68.00	384.20	5.80	4.40	2.00	478.53	7.40	6.00	579.24			
	2	110.16	635.49	1.51	0.86	1.14	213.62	3.79	4.25	302.12			
Lausanne	1	204.80	1,701.60	10.80	6.80	5.40	685.10	14.89	26.08	904.63			
	2	287.24	2,386.62	3.95	2.62	3.97	281.48	10.69	19.75	464.28			
Leeds	1	554.40	16,214.40	16.20	8.60	7.20	2,478.50	86.90	96.80	2,949.60			
	2	647.89	19,646.78	7.59	6.32	3.43	2,745.78	94.49	88.93	3,237.15			
Los Angeles	1	521.60	27,013.20	14.20	12.00	7.40	2,582.44	127.21	162.46	3,137.27			
	2	676.11	41,496.05	4.70	3.81	1.78	2,526.20	188.49	147.95	3,053.15			
New Delhi	1	1,490.40	268,126.80	21.20	12.00	10.00	9,764.81	1,505.77	1,197.61	12,645.40			
	2	1,714.16	309,493.78	6.81	5.35	2.59	7,154.48	1,385.80	1,010.94	9,732.02			
Sao Paolo	1	316.00	7,876.60	12.00	6.80	8.60	1,140.52	49.41	46.61	1,368.60			
	2	411.14	11,299.68	3.41	2.70	1.62	435.66	44.04	42.15	640.02			

Table 5 Detailed results of the case study – Transition duration, number of overlaps, and time required to compute the transition for each transition style. #M = Number of moved labels, Trans. Comp. = Time required to compute the transition, Trans. Durat. = Transition duration.

Scenario	Setting	#M	Naive transitions										DAG-based transitions										Simultaneous transitions									
			#Overlaps			Trans. Comp. [ms]			Trans. Durat. [ms]			#Overlaps			Trans. Comp. [ms]			Trans. Durat. [ms]			#Overlaps			Trans. Comp. [ms]			Trans. Durat. [ms]					
			Max	Avg.	Total	Max	Avg.	Total	Max	Avg.	Total	Max	Avg.	Total	Max	Avg.	Total	Max	Avg.	Total	Max	Avg.	Total	Max	Avg.	Total						
Italy	1	10	2	0.40	2	7.20	2.66	6,499.60	2,200.86	1	0.20	1	9.60	3.38	4,488.30	1,398.12	3	0.60	3	6.80	2.88	2,503.00	999.44									
	2	42	2	0.16	6	7.50	3.26	6,506.20	1,620.26	1	0.05	2	5.80	2.48	5,504.30	1,296.22	3	0.24	9	6.50	2.89	2,502.70	1,079.76									
Lausanne	1	27	2	0.40	2	10.30	5.70	16,489.60	6,498.68	2	0.40	2	13.20	5.90	4,505.20	2,301.50	2	0.40	2	9.30	5.42	2,508.70	1,501.84									
	2	147	7	1.43	53	14.80	7.02	18,496.60	5,273.01	6	0.78	29	14.50	7.41	9,493.80	3,238.79	7	1.19	44	15.60	5.79	2,504.40	1,782.81									
Leeds	1	36	7	1.40	7	13.90	7.10	23,497.00	9,097.04	4	0.80	4	14.30	6.04	5,486.50	2,094.52	5	1.00	5	9.00	4.46	2,493.40	1,095.76									
	2	127	6	1.03	38	11.10	6.46	16,499.50	4,810.41	5	0.59	22	12.40	5.94	7,503.30	2,917.80	4	0.65	24	7.20	4.92	2,507.70	1,783.48									
Los Angeles	1	37	5	1.00	5	13.20	7.64	25,494.40	9,296.16	2	0.40	2	13.10	7.16	4,499.60	2,297.78	2	0.40	2	14.80	8.44	2,494.80	1,493.04									
	2	66	4	0.43	16	8.50	4.70	8,505.80	2,553.35	4	0.30	11	11.00	5.29	4,501.50	1,877.12	3	0.38	14	8.00	5.18	2,502.70	1,525.94									
New Delhi	1	50	3	1.20	6	17.80	9.90	22,481.00	10,796.62	2	0.60	3	15.80	9.28	8,487.50	4,595.52	4	1.40	7	15.80	7.78	2,502.90	2,000.66									
	2	96	3	0.59	22	13.00	7.54	12,498.60	3,539.01	2	0.27	10	13.20	7.42	5,495.70	2,214.52	3	0.46	17	15.50	7.79	2,513.50	1,646.74									
Sao Paolo	1	43	1	0.20	1	11.70	6.50	25,501.00	9,696.42	0	0.00	0	12.50	6.10	8,491.20	3,296.46	3	0.60	3	9.00	5.22	2,500.80	1,497.46									
	2	60	5	0.41	15	10.40	4.86	13,496.00	2,281.82	4	0.24	9	8.20	5.11	8,491.20	1,688.23	4	0.38	14	13.80	4.62	2,503.20	1,201.41									

Reducing False Discoveries in Statistically-Significant Regional-Colocation Mining: A Summary of Results

Subhankar Ghosh ✉

Department of Computer Science & Engineering, University of Minnesota, Minneapolis, MN, USA

Jayant Gupta ✉

Department of Computer Science & Engineering, University of Minnesota, Minneapolis, MN, USA

Arun Sharma ✉

Department of Computer Science & Engineering, University of Minnesota, Minneapolis, MN, USA

Shuai An ✉

Department of Economics, University of Minnesota, Minneapolis, MN, USA

Shashi Shekhar ✉

Department of Computer Science & Engineering, University of Minnesota, Minneapolis, MN, USA

Abstract

Given a set S of spatial feature types, its feature instances, a study area, and a neighbor relationship, the goal is to find pairs \langle a region (r_g), a subset C of S \rangle such that C is a statistically significant regional-colocation pattern in r_g . This problem is important for applications in various domains including ecology, economics, and sociology. The problem is computationally challenging due to the exponential number of regional colocation patterns and candidate regions. Previously, we proposed a miner [8] that finds statistically significant regional colocation patterns. However, the numerous simultaneous statistical inferences raise the risk of false discoveries (also known as the multiple comparisons problem) and carry a high computational cost. We propose a novel algorithm, namely, multiple comparisons regional colocation miner (MultComp-RCM) which uses a Bonferroni correction. Theoretical analysis, experimental evaluation, and case study results show that the proposed method reduces both the false discovery rate and computational cost.

2012 ACM Subject Classification Information systems \rightarrow Data mining; Computing methodologies \rightarrow Spatial and physical reasoning

Keywords and phrases Colocation pattern, Participation index, Multiple comparisons problem, Spatial heterogeneity, Statistical significance

Digital Object Identifier 10.4230/LIPIcs.GIScience.2023.3

Funding This material is based upon work supported by the National Science Foundation under Grants No. 2118285, 2040459, 1901099, and 1916518.

Acknowledgements We also thank Kim Koffolt, Yash Travadi, and the Spatial Computing Research Group for valuable comments and refinements.

1 Introduction

Regional-colocation patterns are (study sub-area R , feature-type subset C) pairs such that instances of feature-types in C often are present in R in close proximity. Given a set S of spatial features (e.g., coffee shops, restaurants), their feature instances, a study area, and a neighbor relationship (e.g., geographic proximity), the goal is to identify pairs \langle region r_g , subset C of S \rangle such that instances of C are statistically significant in that region r_g . Figure 1(a) shows a set of instances input into a regional-colocation miner, consisting of three different spatial feature types, a neighborhood relation between feature instances, and a



© Subhankar Ghosh, Jayant Gupta, Arun Sharma, Shuai An, and Shashi Shekhar; licensed under Creative Commons License CC-BY 4.0

12th International Conference on Geographic Information Science (GIScience 2023).

Editors: Roger Beecham, Jed A. Long, Dianna Smith, Qunshan Zhao, and Sarah Wise; Article No. 3; pp. 3:1–3:18

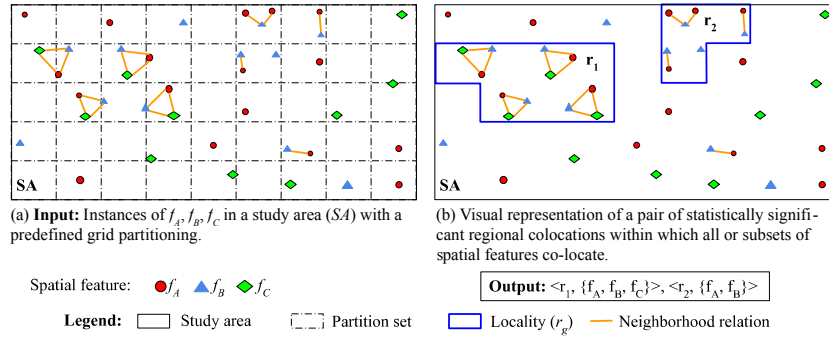
Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

3:2 Reducing False Discoveries in Statistically-Significant Regional-Colocation Mining

space partitioning. Figure 1 (b), shows the set of statistically significant regional colocations identified after significance testing (described in Section 2.2). The output is a pair of regional colocations: r_1 showing a strong regional colocation between all three features (i.e., f_A , f_B , and f_C) and r_2 showing a strong regional-colocation between two features (i.e., f_A and f_B). The rest of the area within the map shows less spatial interaction (low participation index) between these features.

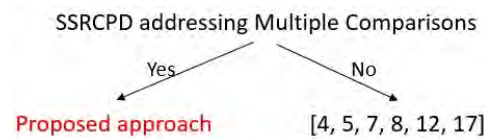


■ **Figure 1** Regions where all or subsets of f_A , f_B and f_C significantly co-locate in the study area.

The problem of mining statistically significant regional-colocation patterns is societally important with applications in retail, public health, ecology, public security, transportation, etc. For example, retail establishments (e.g., fast food chains and coffee shops) often colocate to reach each other’s customers. Thus, finding statistically significant regional colocation patterns among competing retail stores has tremendous value for retail analysis. When identifying colocation patterns in societal domains, it’s important to minimize the chance of false discoveries. A famous historical example was between 1900 and 1904 when urban districts of San Francisco experienced an outbreak of bubonic plague, resulting in 119 deaths. The federal and state authorities falsely identified the victims’ ethnicity as a highly correlated feature to the plague. This false discovery brought an immense adverse impact on San Francisco’s management of the plague. Even when we don’t unfairly stigmatize groups or regions, false discoveries waste money, and resources. Comparing the city’s response to the same plague between 1907 and 1908, where rats were correctly identified as a highly correlated feature and the plague was swiftly contained, the negative impact of false discovery was even more strongly felt [13]. Table 1 provides application domains and use cases.

■ **Table 1** Regional-colocation applications.

Application Do-main	Example
Retail	$\langle \text{China}, \{\text{McDonald's and KFC}\} \rangle, \langle \text{USA}, \{\text{McDonald's and Jimmy John's}\} \rangle$
Public Health	$\langle \text{Ports}, \{\text{Plague and rats}\} \rangle, \langle \text{Middle East}, \{\text{Middle East Respiratory Syndrome (MERS) in 2012 and MERS-CoV}\} \rangle$
Ecology	$\langle \text{Indian/Pacific Ocean}, \{\text{Anemone and Clownfish}\} \rangle, \langle \text{Nile River delta}, \{\text{Nile Crocodile and Egyptian Plover}\} \rangle$
Public Safety	$\langle \text{Region around bars}, \{\text{Assault crimes and drunk driving}\} \rangle$
Transportation	$\langle \text{Near bus depots}, \{\text{High } NO_x \text{ concentrations and buses}\} \rangle$



■ **Figure 2** Comparison with Related Work.

The problem of statistically significant regional-colocation pattern detection (*SSRCPD*) is computationally challenging due to the following reasons: (1) Significance testing in this problem requires considering multiple statistical inferences simultaneously which leads to an increase in Type-I error (i.e false discoveries). (2) There is an exponential number of candidate regional patterns, e.g., the dataset used in the case study (Section 6) consists of 1473 different retail brands and their locations in Minnesota, resulting in 2^{1473} different candidate patterns. (3) Spatial partitioning approach would lead to an infinite number of candidate region subsets.

Figure 2 shows a decision tree that distinguishes our manuscript from previous works, where *SSRCPD* refers to *Statistically significant regional colocation pattern detection*. Earlier work on regional-colocation pattern detection either uses data unaware space partitioning (e.g., Quadtree [4, 12]) or clustering of colocation instances [5, 7]. However, these techniques lack statistical significance testing and depend on input parameters (e.g., participation index threshold) which may vary geographically. Statistically significant global colocation mining was introduced by [1], while statistically significant regional colocation mining was first explored in [8]. In [8] we proposed *SSRCM* which utilizes a subgraph enumeration approach to detect statistically significant regional colocation patterns where the regions would be composed of one or more contiguous atomic partitions (smallest region within which a candidate pattern is statistically significant). This algorithm was expensive because expanding the region within which the pattern was statistically significant required recalculating the *p-value*. Since detecting statistically significant regional colocation patterns requires performing multiple simultaneous statistical inferences, this results in the multiple comparisons problem [14], which risks false discoveries (a.k.a. Type-I errors). The problem results in a rapid increase in the probability of Type-I error as the number of partitions increases. To address the multiple comparisons problem, we propose a robust statistically significant regional colocation miner (MultComp-RCM) using a Bonferroni correction [3]. The proposed approach recommends stricter *p-values* to reduce false discoveries (Type-I errors), thus setting an upper bound on the overall significance level (α , which is 0.05 for a 95% statistical confidence).

Contributions.

- We proposed a new approach Multiple comparisons regional colocation miner (MultComp-RCM) to reduce false positives using a well-established statistical technique for multiple comparisons correction, the Bonferroni test.
- The paper provides a comparative analysis showing that the proposed MultComp-RCM is computationally more efficient than SSRCM.
- The paper describes a sensitivity analysis using synthetic data which shows that MultComp-RCM requires an increasingly smaller number of significance tests and participation index computations for an increasing number of regions.
- We proposed a case study on retail establishments in Minnesota using the Safegraph POI dataset [8]. The proposed method discovers new regional-colocation patterns involving fast food and coffee retailer feature-type subsets in a Minnesota counties study area. We also confirm that the Bonferroni correction in our method reduces false discoveries.

Scope. For simplicity, this paper focuses on regional-colocation patterns consisting of two or three different features. In our case study, we enumerated regions based on a contiguous collection of counties. Nevertheless, this work can be extended to different types of regions (e.g., ports). We also do not consider segregation patterns (negative spatial interaction) or the temporal aspects of the patterns.

Organization. The paper is organized as follows. Section 2 reviews basic concepts and formally defines the problem. In section 3 we briefly review SSRM and describe the proposed approach (MultComp-RCM). Section 4 gives a theoretical analysis of MultComp-RCM. We present the experimental evaluation in Section 5 and a case study in Section 6. Section 7 briefly surveys related work and discussion. Section 8 concludes the paper with future work.

2 Basic Concepts and Problem Definition

First, we review basic concepts related to colocation detection, statistical significance testing, and the multiple comparisons problem. Then, we formally define statistically significant regional colocation pattern detection.

2.1 Colocation detection

In this section, we briefly introduce some taxonomy and the basic concept used to define colocation pattern detection with examples. The basic concepts are as follows:

A **feature instance** is a geo-located spatial entity which is a type of Boolean feature f with a geo-reference point location p (e.g., latitude, longitude), represented as $\langle f, p \rangle$. Multiple instances of a feature are represented as f_i and can be related to other feature instances f_j via a **neighbor relation** \mathcal{R} . For example, geographic proximity is represented as $\mathcal{R}_{f_i, f_j} \leq \theta$, where θ is the neighbor relation threshold. In a **neighbor graph**, we represent features that satisfy such relations as a *node* and their relationship as an *edge*.

A **colocation candidate** C is a set of features defined in the given study area (SA) or a sub-region (r_g) where $r_g \in SA$. For example, Figure 1(a) shows 17 spatial objects of type f_A (circle), 12 spatial objects of type f_B (triangle), and 9 instances of colocation pattern $\{f_A, f_B\}$. An instance of a **colocation** satisfies the neighborhood relation \mathcal{R} and forms a **clique**.

A **participation ratio** (pr) is the ratio of feature instances participating in a relation \mathcal{R} to the total number of instances inside the study region (SA). For a given colocation candidate C and feature f , it is represented as $pr(f, C)$ as shown in Equation 1:

$$pr(f, C) = \frac{\text{participating_instances}(f, C)}{\text{instance}(f)}. \quad (1)$$

For the feature instances shown in Figure 1(a) the participation ratio values for the relation $\{f_A, f_B\}$ are $pr(f_A, \{f_A, f_B\}) = \frac{9}{17}$ and $pr(f_B, \{f_A, f_B\}) = \frac{8}{12}$. Further, the participation ratio within a region (r_g) for a feature f is defined as $pr(f, [r_g, C])$. For example, in Figure 1 $pr(f_A, [r_2, \{f_A, f_B\}])$ and $pr(f_B, [r_2, \{f_A, f_B\}])$ in region r_2 and has the value $\frac{4}{4}$ and $\frac{3}{4}$ respectively.

A **participation index** (pi) is the minimal participation ratio of all feature types in a colocation candidate as described in Equation 2:

$$pi(C) = \min_{f \in C} (pr(f, C)). \quad (2)$$

The participation index quantifies the spatial interaction within features. Figure 1(a) shows participation index of features f_A, f_B which can be represented as $pi(\{f_A, f_B\})$ which is $min(\frac{9}{17}, \frac{8}{12})$ or $\frac{8}{17}$. A **regional participation index** is the minimal participation ratio of all feature types in the colocation candidate C within region r_g as shown below

$$pi([r_g, C]) = \min_{f \in C}(pr(f, [r_g, C])) \quad (3)$$

For instance in Figure 1, $pi([r_2, \{f_A, f_B\}]) = min(\frac{4}{4}, \frac{3}{4}) = \frac{3}{4}$.

Colocation patterns [16] is the set of prevalent colocation candidates (based on a prevalence measure, e.g. pi), i.e., candidates comprised of features having a high positive spatial interaction. A **regional-colocation pattern** [12] is a paired region (r_g) and colocation pattern (C), i.e., $\langle r_g, C \rangle$ where the features in pattern C have a high positive spatial interaction in r_g .

2.2 Statistical Significance in Colocation Detection

A statistically significant colocation determines whether an assigned positive spatial interaction between features is statistically significant or could have been observed if the features were in complete spatial randomness (CSR). Other properties in CSR are as follows:

- Every feature instance has an equal probability of existing at any point in the study area.
- The locations of any feature instances in the study area are independent of each other.

A **null hypothesis** (H_0) is a statement of “no effect” or “no difference”. In our problem, the null hypothesis represents the scenario under which there is no spatial interaction between the features in the dataset, i.e., their existence is completely independent of each other.

An **alternative hypothesis** (H_a) is a statement that is tested against a null hypothesis. In our problem, an alternative hypothesis represents the scenario under which there is a positive spatial interaction between the features in the dataset in a region of interest.

A **Type-I error** refers to the erroneous rejection of an actually true null hypothesis (or a false positive). In our problem, this would refer to incorrectly assigning a candidate regional-colocation pattern as statistically significant, even though there is a high probability of this pattern being found in CSR or H_0 .

A **Type-II error** refers to the failure to reject a null hypothesis (H_0) that is actually false (or a false negative). This would translate into incorrectly assigning a candidate regional-colocation pattern as not statistically significant.

A **point distribution** is a collection of geo-distributed points referring to an event (e.g., road accident) in a spatial domain. A **point process** (PP) is a statistical process that defines the probability distribution of a point over a region. Point processes are essential for defining the null or alternative hypothesis for our statistical significance test.

A **Poisson point process** is defined in a generalized space S_P with intensity Λ having the following properties:

1. The number of points in a bounded Borel set (bounded sets that can be constructed from open or closed sets by repeatedly taking countable unions and intersections) $B \subset S_P$ is a Poisson random variable with mean $\Lambda(B)$.
2. The number of points in n disjoint Borel sets forms n independent random variables. This property results in independent scattering or complete independence.

Null hypothesis generation.

- For an identical distribution, we generate an equal number of instances of each feature in every partition using summary statistics of the constituent features of the pattern. This ensures that the null hypotheses datasets (although in CSR) closely model the observed dataset in each atomic partition.

Most techniques to address this problem require a stricter significance threshold for individual comparisons to compensate for the number of inferences being made. A stated confidence level generally applies to individual tests. It is often desirable to have a confidence level for a whole family of simultaneous tests.

The **Bonferroni correction** [3] is a method to address the multiple comparisons problem and the simplest method for reducing Type-I errors. It is a conservative method with a greater risk of failure to reject a false null hypothesis, thus resulting in Type-II errors.

2.3 Formal problem formulation

The problem of statistically significant regional-colocation pattern detection is as follows:

Input:

1. A set (F) of spatial-features
2. N geo-located spatial feature instances.
3. A study area S_A composed of space partitions (e.g., counties).
4. A statistical significance level α .
5. A neighbor relationship (\mathcal{R}).

Output: Statistically significant regional-colocation patterns, $\langle r_g, C \rangle$ where $C \subset F$.

Objective: Reducing Type-I error (false positives).

Constraints: Higher statistical confidence of output patterns.

Reasoning behind problem output. Testing for statistical significance on regional-colocation outputs ensures that spurious patterns aren't detected from the dataset. Otherwise, regions may be enumerated due to a high density of feature instances or spatial auto-correlation. In addition, significance testing for the union of many partitions leads to multiple statistical inferences. Due to the union of partitions, the probability of finding chance patterns within the bigger region (i.e., the union of partitions) is higher; this phenomenon is not accounted for by the p -value threshold for a single partition. This leads to the multiple comparisons problem, resulting in a higher false discovery rate. In application domains related to regional-colocation pattern detection, reducing Type-I errors (false positives) takes higher priority over reducing Type-II errors (false negatives). These Type-II errors might result in missing the detection of certain patterns which might have a lower p -value.

In this situation, checking for a particular α level in the individual statistical inferences is insufficient. We also need to control the family-wise error rate which represents the probability of making one or more false discoveries (Type-I errors) [14]. We use the Bonferroni correction in MultComp-RCM to tackle this problem arising from multiple hypothesis tests. This conservative approach ensures that the pattern output has high statistical confidence while ignoring patterns that might have comparatively lower confidence, which is our primary objective. Another benefit of this method is the computational efficiency due to the smaller number of significance tests and participation index computations required as compared to the baseline [8]. The Bonferroni correction proposes stricter p -value thresholds which might be a bottleneck for large scale applications, such as when dealing with hundreds of atomic partitions. This may also lead to a higher possibility of false negatives (Type-II errors).

3 Methodology

To keep the paper self-contained, we first briefly review the SSRCM, our previous statistically significant regional-colocation miner [8], and a sub-routine on significance testing. We then describe the proposed approach in Section 3.2 and provide an example highlighting the computational cost savings of the new approach.

3.1 Statistically Significant Regional-Colocation Miner

Key idea. In [8], we started by considering partitions with at least 3 instances of each feature which comprise the regional-colocation pattern. This ensures that the features constituting the pattern all have a considerable presence in the enumerated partitions. We then use the regional statistical significance test as described in Algorithm 1 to determine the atomic footprints of the pattern, i.e., statistically significant pattern within individual partitions. While computing the participation index, we limit our neighborhood to an empirically determined distance (d) to mine meaningful collocated features.

■ **Algorithm 1** Significance testing.

Input:

- A spatial dataset S consisting of features $\{f_A, f_B, \dots\}$
- A study area (S_A) and an atomic partition $r_g \subset S_A$
- Statistical significance level α
- A candidate colocation pattern C
- A set of R Null hypotheses (NH_θ) data each modelled as colocation C in atomic partition r_g
- Distance d for participation index (pi) calculation

Output:

1. $\langle r_g, C \rangle$ is significant or not
2. $p\text{-value}_C^{r_g}$

1: **procedure** SIGNIFICANCE TESTING
2: Statistically significant result $SSR_C^{r_g} \leftarrow \text{False}$
3: Counter $R^{\geq pi_{obs}} \leftarrow 0$
4: Calculate pi_{obs} for C at d in r_g
5: **for** $i \in [1, R]$ **do**
6: Calculate the $pi_{\theta, i}$ of C at d in the i^{th} NH_θ
7: **if** $pi_{\theta, i} \geq pi_{obs}$ **then**
8: $R^{\geq pi_{obs}} \leftarrow R^{\geq pi_{obs}} + 1$
9: $p\text{-value}_C^{r_g} = \frac{R^{\geq pi_{obs}} + 1}{R + 1}$
10: **if** $p\text{-value}_C^{r_g} \leq \alpha$ **then**
11: $SSR_C^{r_g} \leftarrow \text{True}$ ▷ (i.e., $\langle r_g, C \rangle$ is statistically significant)
12: **else**
13: $SSR_C^{r_g} \leftarrow \text{False}$ ▷ (i.e., $\langle r_g, C \rangle$ is not statistically significant)
14: **return** $SSR_C^{r_g}, p\text{-value}_C^{r_g}$

For finding the union of partitions, we first form an undirected unweighted graph ($G = (V, E)$) where each vertex (V) refers to a partition within which the pattern is statistically significant, and an edge (E) between two V s represents a shared boundary between them. The graph representation allows the use of graph traversal algorithms (e.g., DFS) to find statistically significant regions which are the union of partitions in V .

We note that the union of two atomic footprints within which a candidate regional colocation pattern is statistically significant does not imply that the resultant footprint is a significant regional-colocation pattern. Thus, we need to recompute the pi for the candidate pattern in the new region and perform the significance test again. As we progress along the edges of G , the final output is a larger region composed of contiguous atomic partitions such that the candidate pattern is statistically significant, both within the atomic partitions as well as in the region formed by the union of the output atomic footprints. This is represented by the **largest connected component**. Algorithm 2 provides the pseudo-code of $SSRCM$ to find statistically significant regional colocations.

Algorithm 2 Statistically Significant Regional-Colocation Miner (SSCRM).

Input:

- A Spatial dataset S consisting of features $\{f_A, f_B, \dots\}$
- A study area (S_A) and a space partitioning R_g
- Statistical significance level α
- Maximum pattern size N
- Lower bound LB (in meters)
- Upper bound UB (in meters)

Output:

1. List of statistically significant regional colocation patterns [$\langle r_g, C \rangle$]

Variables:

Distance between feature instances d

```

1: procedure STATISTICALLY SIGNIFICANT REGIONAL-COLOLOCATION MINER
2:   for each:  $f_k$  in  $\{f_A, f_B, \dots\}$  do
3:     Generate  $R$  null hypotheses ( $NH_\emptyset$ ) using summary statistics in each  $r_g \in R_g$ .
4:   for each: candidate pattern  $C_m \in \{C_1, C_2, \dots, C_M\}$  do
5:     for distance  $d \in [LB, LB + 10, \dots, UB]$  do
6:       for each:  $r_g \in R_g$  do
7:          $SSR_{C_m}^{r_g}, p\text{-value} \leftarrow$  Significance Testing( $S, r_g, \alpha, C_m, NH_\emptyset, d$ )
8:         if  $SSR_{C_m}^{r_g}$  is True then
9:           Insert  $r_g$  in significant atomic partitions list
10:        Compose Neighborhood graph ( $G$ ) from significant atomic partitions list
11:         $r_g^{final} \leftarrow r_g^{maxPI}$   $\triangleright$  atomic partition in  $G$  with highest  $pi$ 
12:        for each:  $r_g \in$  Depth First Graph Traversal of  $G$ 
13:          from vertices adjacent to  $r_g^{maxPI}$  do  $\triangleright r_g \neq r_g^{maxPI}$ 
14:             $r_g^{temp} \leftarrow r_g^{final} \cup r_g$ 
15:             $SSR_{C_m}^{r_g}, p\text{-value} \leftarrow$  Significance Testing( $S, r_g^{temp}, \alpha, C_m, NH_\emptyset, d$ )
16:            if  $SSR_{C_m}^{r_g}$  is True then
17:               $r_g^{final} \leftarrow r_g^{temp}$ 
18:            Add  $\langle r_g^{final}, C_m \rangle$  to [ $\langle r_g, C \rangle$ ]
19:   return [ $\langle r_g, C \rangle$ ]

```

3.2 Multiple Comparisons Regional Colocation Miner (MultComp-RCM)

Key Idea. We observe that the baseline *SSRCM* computes a significance test for every union of statistically significant partitions, resulting in many participation index (pi) computations and significance tests. We address this by using a Bonferroni correction, which selects atomic partitions conservatively, increasing the chances that their union is also statistically significant. The Bonferroni correction reduces the need to perform a regional statistical significance test for each union operation.

A Bonferroni correction is used when several independent statistical inferences are being performed simultaneously. Although a given significance threshold (α) on the p -value may be appropriate for an individual test, it is not sufficient for the set of all comparisons. To reduce many false positives, the α needs to be lowered to account for the number of comparisons performed. The Bonferroni correction sets the statistical significance threshold for the entire set of n comparisons to α/n or, equivalently, by multiplying the p -value by n , and then applying the standard threshold α . This conservative correction works even under the most extreme circumstances (e.g., when all n tests are independent of one another).

In the proposed approach we check for statistical significance in each input partition. Then, we perform a graph traversal starting from the atomic partition with the highest pi value for the candidate regional-colocation pattern. Then, instead of recomputing the pi and testing the candidate pattern in the new bigger region (composed of atomic partitions) for

3:10 Reducing False Discoveries in Statistically-Significant Regional-Colocation Mining

statistical significance, we perform a Bonferroni correction. Thus if we were initially checking for a threshold level of 0.05, then for the union of two partitions, we would be checking for a threshold level of $0.05/2$ in each atomic partition. This conservative threshold reduces Type-I error by returning regions with much higher statistical confidence. The union of the atomic partitions is sequential and every atomic partition must satisfy the adjusted p-value threshold to be considered for the union.

Algorithm 3 provides a snippet of MultComp-RCM showing the use of the Bonferroni correction. Lines 13-18 show the new steps in the refined approach.

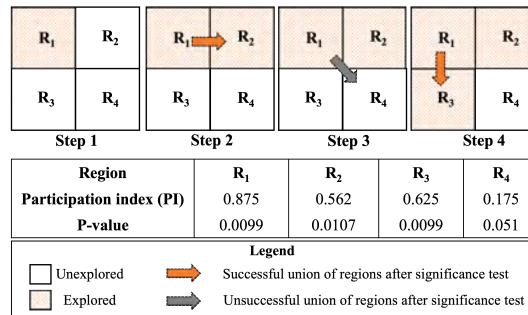
■ **Algorithm 3** MultComp-RCM snippet.

```

1: procedure MULTCOMP-RCM
2:   ⋮
12:   $n \leftarrow 1$  ▷ Number of atomic partitions in the region.
13:  for each:  $r_g \in$  Depth First Graph Traversal of  $G$ 
14:    from vertices adjacent to  $r_g^{maxPI}$  do ▷  $r_g \neq r_g^{maxPI}$ 
15:      Subgraph ( $SG$ )  $\leftarrow r_g^{final} \cup r_g$ 
16:      flag  $\leftarrow 1$ 
17:      flag = BONF_CHECK(flag,  $SG$ ,  $n$ )
18:      if flag == 1 then
19:        Update  $r_g^{final} \leftarrow SG$ 
20:         $n \leftarrow n + 1$ 
21:      Add  $\langle r_g^{final}, C_m \rangle$  to  $\langle r_g^C, C \rangle$ 
22:    ⋮
23: procedure BONF_CHECK(flag,  $SG$ ,  $n$ )
24:   $p\text{-value}_{threshold} \leftarrow \alpha / (n + 1)$ 
25:  for each:  $node \in SG$  do
26:    if  $p$ -value of pattern  $C_m$  in  $node \not\leq p\text{-value}_{threshold}$  then
27:      flag  $\leftarrow 0$ 
28:  return flag

```

Figure 4 shows an execution trace of merging 4 neighboring partitions. Each region has a participation index and a p -value computed individually. Then, Steps 1-4 show the process of combining these partitions based on either additional statistical significance tests and participation index computations (for SSRM) or using a tighter p -value threshold for MultComp-RCM. Table 2 compares the number of computations for the two approaches and clearly shows the lower computational requirements of MultComp-RCM. When performing the union of two regions using MultComp-RCM, the new threshold as per the Bonferroni correction is applied to each of the two regions (as in procedure BONF_CHECK in Algorithm 3) for a successful union.



■ **Figure 4** Execution trace of SSRM and MultComp-RCM.

■ **Table 2** Comparing the cumulative number of statistical significance tests ($C\#$), participation index computation (pi cal.), and p-value thresholds (p -val th.) between $SSRCM$ (denoted as S) and $MultComp$ -RCM (denoted as R).

Steps	$C\#S$	$C\#R$	pi cal. S	pi cal. R	p -val. th. S	p -val. th. R
0	4	4	4	4	0.05	0.05
1	5	4	5	4	0.05	0.05
2	6	4	6	4	0.05	0.025
3	7	4	7	4	0.05	0.0167
4	8	4	8	4	0.05	0.0125

4 Theoretical Analysis

► **Lemma 1.** *MultComp – RCM has lower or equal Type-I error than SSRCM.*

Proof. Algorithm 1 called by Algorithm 2 and 3 in line 7 extracts atomic partitions within which a regional-colocation pattern is statistically significant.

The Bonferroni correction in procedure BONF_CHECK in Algorithm 3 controls the experiment-wide false positive rate (π) by specifying the significance level (α) for each test, where a test is significant if p -value $\leq \alpha$. The probability of no Type I error (false positives) in n independent tests is $(1 - \alpha)^n$, if each test is at level α . Therefore, the probability of at least one false positive π is $1 - (1 - \alpha)^n$. For an experiment-wide false positive rate of π , the α for each test should be $\alpha = 1 - (1 - \pi)^{1/n}$. Using binomial approximation, $(1 - \alpha)^n \simeq 1 - n\alpha$, which gives $\alpha = \pi/n$. For an experiment-wide false positive value $\pi = 0.05$, the α (false positive rate for each test) should be less than π , i.e. $\alpha \leq \pi$. Therefore each region and sub-region output by MultComp-RCM has lower Type-I and a precision close to 1. ◀

► **Lemma 2.** *MultComp-RCM has lower or equal computational cost than SSRCM for all observed data, where Bonferroni-revised p-values eliminate lower confidence candidates considered by the original p-value, i.e. $Cost_{MultComp-RCM} \leq Cost_{SSRCM}$.*

Proof. Let $C_{pi}(d)$ be the complexity of participation index (pi) computation for a specific region (dependent on the data d). Let $C_{st}(pi_{obs}, pi_{null}, d)$ be the complexity of significance testing for a specific region (dependent on the pi in observed data d and the null hypothesis). Assume N_1 is the number of space partitions/regions in the dataset, N_2 is the number of space partitions extracted from Algorithm 1, and $N_2 \leq N_1$. Further, assume d_1 is the initial dataset and d_2 is the dataset in each iteration in the SSRCM. Then, the cost of $SSRCM$ is

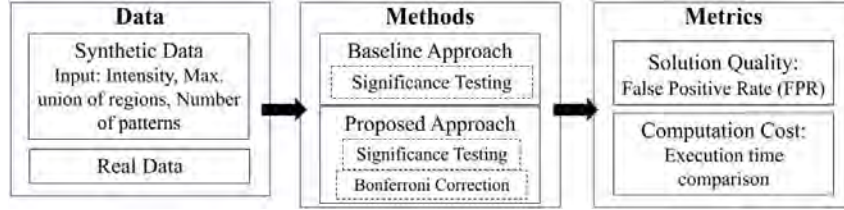
$$N_1(C_{pi}(d_1) + C_{st}(pi_{obs}, pi_{null}, d_1)) + N_2(C_{pi}(d_2) + C_{st}(pi_{obs}, pi_{null}, d_2)) \quad (6)$$

By contrast, the cost of the proposed MultComp-RCM approach is only $N_1(C_{pi}(d_1) + C_{st}(pi_{obs}, pi_{null}, d_1)) + N_2$. Here, N_2 represents the number of significant partitions for which the p-value needs a comparison against the threshold obtained from Bonferroni correction. ◀

5 Experimental Evaluation

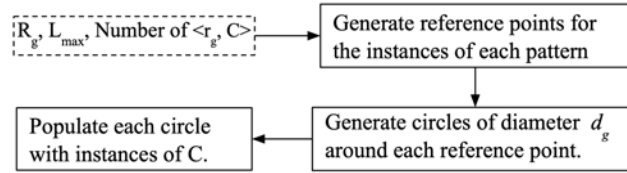
We had three goals for the experiments: (1) To compare the time taken by $SSRCM$ and MultComp-RCM with varying numbers of regional-colocation instances, varying number of atomic partitions, and change in the number of feature instances. (2) To compare the number of significance tests, pi calculations for a varying number of regions. (3) To compare solution quality between $SSRCM$ and MultComp-RCM.

Experiment design. Figure 5 shows the overall validation framework. The metric for comparing the solution quality of *SSRCM* with MultComp-RCM was the false positive rate (FPR), while the runtime comparisons were based on the execution time (in seconds) of the individual algorithms. The experiments were done on both real (Safegraph POI) and synthetic data to perform both comparative and sensitivity analysis.



■ **Figure 5** Overall validation framework.

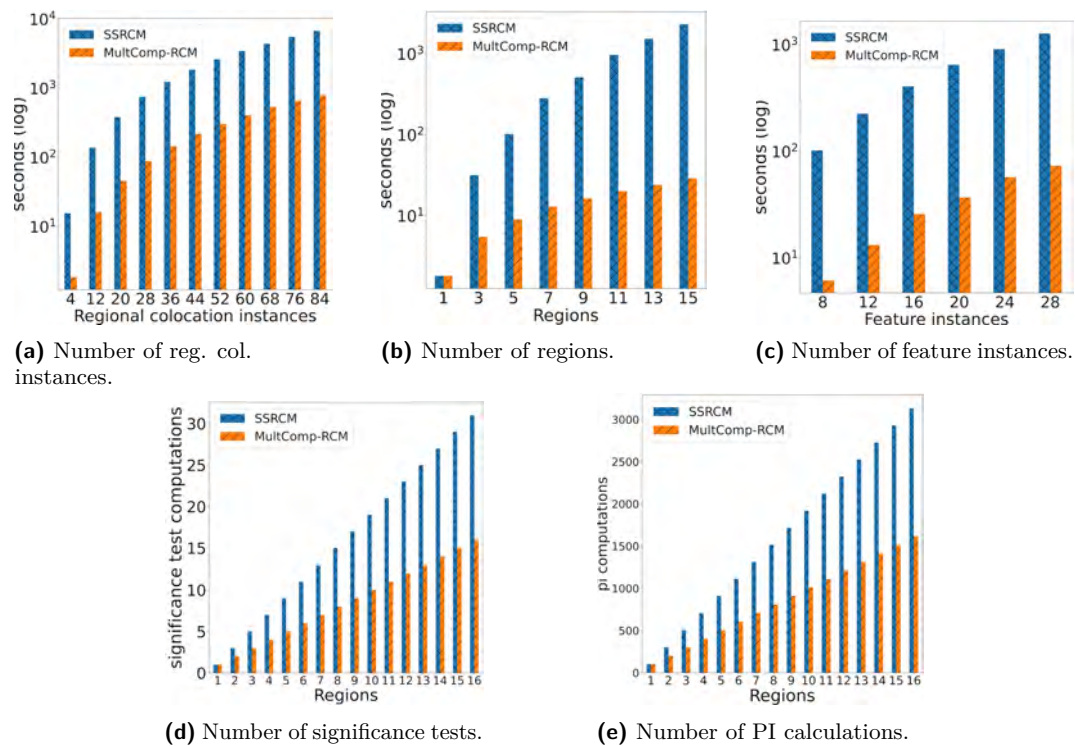
Synthetic data generation. We began with a space partitioning (R_g), a maximum union (or traversal) of regions (L_{max}), and a number of regional-colocation patterns i.e., pairs of $\langle r_g, C \rangle$. We then generated reference points within the partitions using the Poisson point process. At each reference point, we generated circles of diameter d_g which was determined empirically for each region in R_g in the observed dataset. The diameter signifies the smallest distance between features in a colocation C at which they become statistically significant regional colocations. We populated each circle with instances of C . We note that the circles were only used to place collocated instances in a region and were not separate partitioning. Figure 6 shows the process of synthetic data generation.



■ **Figure 6** Synthetic data generation process.

Comparative Analysis. Figure 7a shows the time taken (in log scale) for different regional-colocation instances. For this experiment, we varied the number of regional-colocation instances in each atomic partition from 4 to 84 while keeping other parameters (like the number of regions) constant and record the execution time of both algorithms. Figure 7b compares the execution time with a varying number of atomic partitions (or regions) while keeping the number of regional-colocation instances in each partition constant. Figure 7c shows the time taken with a varying number of feature instances (which constitute the regional-colocation pattern) in each region while keeping the number of regions constant. In all experiments, *MultComp-RCM* is much faster than the baseline *SSRCM*. These results are consistent with Lemma 2, which says that $Cost_{MultComp-RCM} \leq Cost_{SSRCM}$.

Sensitivity Analysis. Figure 7d shows the number of significance tests performed by both algorithms with varying number of regions, while keeping the number of regional-colocation instances constant in each partition. Figure 7e shows the number of participation index



■ **Figure 7** MultComp-RCM outperforms SSRCM [8].

computations performed with varying number of regions with the same constant parameters as above. In both cases, the proposed MultComp-RCM requires lesser number of significance tests and participation index computations for an increasing number of regions.

Solution Quality. We performed controlled experiments on synthetic datasets to compare the solution quality of *MultComp-RCM* with *SSRCM*. Metric for comparison was the false positive rate (FPR).

$FPR = \frac{FP}{FP+TN}$, where FP is the number of false positives, and TN is the number of true negatives. Table 3 shows the experiment results. As shown MultComp-RCM exhibits a lower rate for false pattern discovery than *SSRCM*. This is mainly because MultComp-RCM eliminates regions which barely pass the atomic significance test (borderline statistical confidence) in Algorithm 1, which *SSRCM* fails to reject in the final output.

■ **Table 3** MultComp-RCM generates less false positives.

Pattern	SSRCM False Positive Rate	MultComp-RCM False Positive Rate
A, B, C	0.15	0.03
A, B	0.17	0.01
B, C	0.14	0.01
A, C	0.19	0.04

6 Case Study

We extended our previous case study [8] to show the effectiveness of the proposed approach.

Dataset. We used data from SafeGraph, a mobility data vendor who provides anonymized aggregated location data to researchers studying the effects of COVID-19 on citizen mobility patterns towards numerous Points Of Interest (POIs). The dataset consists of 1473 retail brands in Minnesota. Experiments were performed on colocation patterns consisting of two (e.g., Jimmy John’s, McDonald’s) or three (e.g., Jimmy John’s, McDonald’s, Subway) features. Our null hypothesis generation followed the procedure described in Section 2.2.

Case Study Results. The pattern $C := \{Jimmy\ John's, McDonald's, Subway\}$ was found to be statistically significant when the distance between feature instances was about 1400 meters. The regional footprint was the union of “Dakota” and “Hennepin” Counties. The pi values in the counties were 0.34 and 0.45 respectively. The p -value for the pattern within the counties were 0.02 and 0.01, satisfying the p -value threshold of $\frac{0.05}{2}$ as per the Bonferroni correction for the two partitions. A few additional significant patterns are shown in Table 4 (values rounded to two decimal places).

■ **Table 4** Regional-colocation patterns found to be statistically-significant at distance d .

Colocated features	Counties (<i>participationindex</i> , p – value)	d
{Caribou coffee, Starbucks}	Hennepin (0.34, 0.01)	200 m
{Caribou coffee, Starbucks}	Carver (0.5, 0.02), Hennepin (0.51, 0.01), Washington (0.41, 0.01)	400 m
{Caribou coffee, Starbucks, Dunn Bros}	Hennepin (0.52, 0.01)	1900 m
{Caribou coffee, Starbucks, Dunn Bros}	Hennepin (0.72, 0.01), Washington (0.36, 0.02)	3000 m
{Jimmy John’s, McDonald’s}	Hennepin (0.39, 0.01)	500 m
{Jimmy John’s, McDonald’s}	Dakota (0.36, 0.02), Hennepin (0.51, 0.01)	700 m
{Jimmy John’s, McDonald’s, Subway}	Dakota (0.34, 0.02), Hennepin (0.45, 0.01)	1400 m
{Jimmy John’s, McDonald’s, Subway}	Dakota (0.47, 0.02), Hennepin (0.57, 0.01), Washington (0.43, 0.02)	1500 m

In our previous paper [8], we compared SSRCM with the Quad and QGFR algorithms [12] whose data-aware space partitioning approach is based on the minimum orthogonal bounding rectangle (MOBR). We found that the MOBR-based approach with a participation index threshold of 0.6 produced 3368 potential localities for the pattern $\{r_g, [\text{Caribou Coffee, Starbucks}]\}$. With a confidence level of 95%, MOBR-based approach resulted in 2917 significant and 451 non-significant patterns. Hence, a regional-colocation miner without statistical significance may enumerate output regions where colocations occurred by chance.

7 Related Work and Discussion

Related Work. The concept of colocation was introduced by Shekhar et al. [16]. Huang et al. [10] provided extensive experiments and rigorous discussions regarding the topic and the participation index as a prevalence measure between constituent features. Later, Barua et al. [1] introduced statistical significance testing in global colocation and segregation pattern detection to avoid enumeration of chance patterns in the dataset for both aggregation and segregation patterns but did not mention patterns that are regional (or local). Regional colocation with minimum orthogonal bounding rectangle (MOBR) based approach was studied

by Li et al. [12] while [17] and [4] focused on shapes and zonal patterns, respectively. These methods utilized a threshold on the participation index (pi) without statistical significance testing, leading to the detection of spurious patterns (as discussed in [8]). We [8] recently proposed a subgraph-based approach that incorporates statistical significance in detecting regional colocation patterns. This approach reduced the number of spurious patterns detected by previous methods. However, due to a large number of simultaneous statistical inferences, an increase in false discoveries is also observed. Besides, other patterns [15] and several statistical significance and false discovery reduction techniques have been studied in association rule mining [18, 6]. However, these approaches do not address the inherent variability in spatial data (i.e., different summary statistics of features in each atomic partition). To find subgroups of items, which are generally observed to be statistically significant associations, they compare a quality measure (which assigns to each itemset a numeric value) on the subgroup against that in a statistical model (which corresponds to the null hypothesis). These null hypotheses for significance testing are uniform and do not address spatial variability. Thus these approaches are not directly applicable to regional colocation patterns (more details in Appendix A).

8 Conclusion and Future Work

In this paper, we refined the problem of the statistically significant regional-colocation pattern (*SSCRP*). We proposed a robust *MultComp-RCM* approach that reduces the number of false positives using a Bonferroni correction. We theoretically show that *MultComp-RCM* has a lower or equal Type-I error and computational cost than *SSRCM* along with experimental results. We extended the previous case study on retail establishments in Minnesota using the proposed approach showing a contrast between significant and non-significant patterns.

Future Work. We plan to explore other methods to reduce Type-I errors (false positives) while also addressing Type-II errors (false negatives) arising from the conservative Bonferroni correction approach and further add temporal dimension to these patterns.

References

- 1 Sajib Barua and Jörg Sander. Mining statistically significant co-location and segregation patterns. *IEEE TKDE*, 26(5):1185–1199, 2013.
- 2 Julian Besag and Peter J Diggle. Simple monte carlo tests for spatial pattern. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 26(3):327–333, 1977.
- 3 Carlo Bonferroni. Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, 8:3–62, 1936.
- 4 M Celik et al. Zonal co-location pattern discovery with dynamic parameters. *ICDM*, 2007.
- 5 Min Deng et al. Multi-level method for discovery of regional co-location patterns. *IJGIS*, 2017.
- 6 Wouter Duivesteijn and Arno Knobbe. Exploiting false discoveries—statistical validation of patterns and quality measures in subgroup discovery. In *2011 IEEE 11th International Conference on Data Mining*, pages 151–160. IEEE, 2011.
- 7 Christoph F. Eick, Rachana Parmar, et al. Finding regional co-location patterns for sets of continuous variables in spatial datasets. In *SIGSPATIAL*, 2008.
- 8 Subhankar et. al. Towards geographically robust statistically significant regional colocation pattern detection. In *Proceedings of the 5th ACM SIGSPATIAL GeoSIM*, pages 11–20, 2022.
- 9 Yan Li et al. Cscd: Towards spatially resolving the heterogeneous landscape of mxif oncology data. In *Proceedings of the 10th ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data, BigSpatial '22*, pages 36–46, New York, NY, USA, 2022. ACM.

- 10 Yan Huang et al. Discovering colocation patterns from spatial data sets: a general approach. *IEEE TKDE*, 16(12):1472–1485, 2004.
- 11 Janine Illian, Antti Penttinen, et al. *Statistical analysis and modelling of spatial point patterns*, volume 70. John Wiley & Sons, 2008.
- 12 Yan Li and Shashi Shekhar. Local co-location pattern detection: a summary of results. In *GIScience*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2018.
- 13 Guenter B Risse. “A long pull, a strong pull, and all together”: San francisco and bubonic plague, 1907-1908. *Bulletin of the History of Medicine*, 66(2):260–286, 1992.
- 14 G Rupert Jr et al. *Simultaneous statistical inference*. Springer Science & Business Media, 2012.
- 15 Arun Sharma, Jayant Gupta, and Subhankar Ghosh. Towards a tighter bound on possible-rendezvous areas: preliminary results. In *Proceedings of the 30th International Conference on Advances in Geographic Information Systems*, pages 1–11, 2022.
- 16 Shashi Shekhar and Yan Huang. Discovering spatial co-location patterns: A summary of results. In *Intl. symposium on spatial and temporal databases*, pages 236–256. Springer, 2001.
- 17 Song Wang et al. Regional co-locations of arbitrary shapes. In *SSTD*, 2013.
- 18 Geoffrey I Webb. Discovering significant patterns. *Machine learning*, 68(1):1–33, 2007.
- 19 David WS Wong. The modifiable areal unit problem (maup). In *WorldMinds: geographical perspectives on 100 problems: commemorating the 100th anniversary of the association of American geographers 1904–2004*, pages 571–575. Springer, 2004.

In this appendix, we address the following questions:

A Why can’t we use existing false discovery reduction techniques from local pattern mining?

Existing techniques for reducing false discoveries in local pattern mining cannot be applied to this problem, because of spatial variability (i.e. constituent features of a regional colocation pattern might have different summary statistics in different atomic partitions). Webb [18] proposed a holdout approach where one divides the data into exploratory and holdout sets. Patterns are generated using the exploratory data, while statistical tests are performed on the generated patterns using the holdout data. This technique may apply to atomic partitions with a large presence of constituent features (e.g., partition T41 in Figure 8). However, it would be counterproductive in partitions where the number of feature instances is very low (e.g., partition T42 in Fig. 8). In such partitions splitting the data points into exploratory and holdout sets would result in very few instances for the pattern detection process.

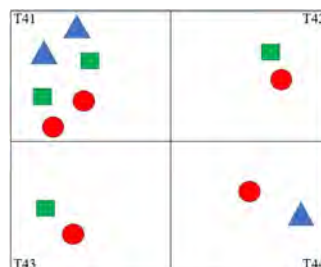


Figure 8 Feature instances exhibit spatial variability within atomic partitions.

B Why can't this problem be cast as a modified version of frequent itemset mining?

In frequent itemset mining, the task is to find subgroups of items that often occur together in a transaction, e.g., laptop and antivirus software. Previous works have been done on addressing false discoveries in this problem [6]. Such approaches assign the association in the mined subgroup as the alternate hypothesis while the null hypothesis is formulated using a randomized baseline subset. Thus these approaches do not address the independent relationship between hypotheses in different spatial partitions in our problem. As noted earlier, in regional colocation pattern detection, different features might have different summary statistics in different atomic partitions. To model the complete spatial randomness of these features, we generate the null hypotheses in each atomic partition as per the summary statistics of the said features in that specific partition. Thus the null hypothesis generated for the features in one atomic partition is independent of the null hypothesis in other atomic partitions. Therefore, the problem of regional colocation pattern detection cannot be considered a modified version of subgroup discovery in frequent itemset mining.

C How does spatial colocation mining differ from association rule mining?

Data mining techniques have been widely developed to solve challenging problems in various domains. Yet, the underlying assumption of these algorithms does not address the problem of spatial variability. This leads to the detection of spurious patterns in spatial data, also known as the modifiable aerial unit problem (MAUP [19]). Colocation pattern detection resembles association rule mining, but the absence of transactions in colocation mining means techniques in association rule mining cannot be used directly to mine colocation patterns.

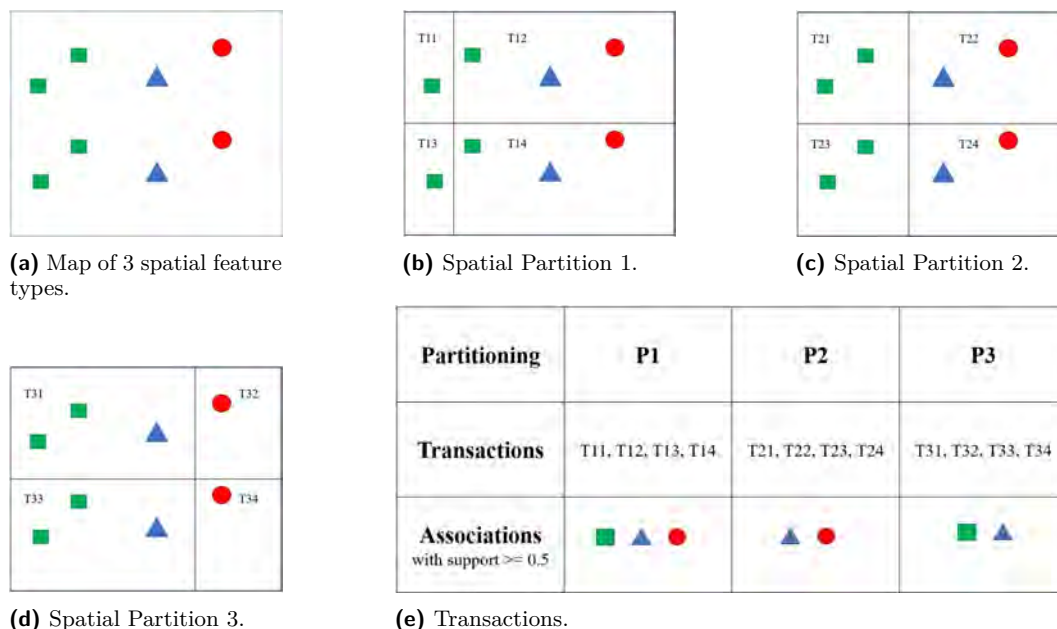
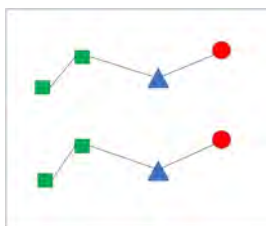






Figure 9 Association rule mining [9] returning different results depending on the spatial partition.

3:18 Reducing False Discoveries in Statistically-Significant Regional-Colocation Mining

Transactions in association rule mining refer to groups of items purchased together. An itemset's support is the fraction of transactions that contain the itemset. Itemsets greater than a user-specified support value yield to the association rule. In spatial data mining, the choice of partition affects the transaction. For example, Figure 9a below shows a dataset with 3 feature types, i.e. <squares>, <triangles>, <circles>. In partition P1 (Figure 9b) <squares, triangles, circles> is a transaction, while in partitions P2 (Figure 9c) and P3 (Figure 9d) <triangles, circles> and <squares, triangles> are the transactions respectively. This is known as the MAUP problem. In colocation pattern detection this is addressed using a neighborhood graph as shown in Figure 10a. A user-defined neighbor relationship R is used to find subsets of features in close geographic proximity. Thus the colocation miner provides a transaction-free approach to mine prevalent patterns.



(a) Neighbor graph based on relation R .

Colocation with Participation Index (PI) ≥ 0.5	PI  	PI  
	= $\min(1/2, 1) = 1/2$	= $\min(1, 1) = 1$

(b) PI of candidate patterns.

■ **Figure 10** Colocation pattern detection [8].

Genetic Programming for Computationally Efficient Land Use Allocation Optimization

Moritz J. Hildemann¹ ✉ 

Institute for Geoinformatics, University of Münster, Germany

Alan T. Murray ✉ 

Department of Geography, University of California at Santa Barbara, CA, USA

Judith A. Verstegen ✉ 

Department of Human Geography and Spatial Planning, Utrecht University, The Netherlands

Abstract

Land use allocation optimization is essential to identify ideal landscape compositions for the future. However, due to the solution encoding, standard land use allocation algorithms cannot cope with large land use allocation problems. Solutions are encoded as sequences of elements, in which each element represents a land unit or a group of land units. As a consequence, computation times increase with every additional land unit. We present an alternative solution encoding: functions describing a variable in space. Function encoding yields the potential to evolve solutions detached from individual land units and evolve fields representing the landscape as a single object. In this study, we use a genetic programming algorithm to evolve functions representing continuous fields, which we then map to nominal land use maps. We compare the scalability of the new approach with the scalability of two state-of-the-art algorithms with standard encoding. We perform the benchmark on one raster and one vector land use allocation problem with multiple objectives and constraints, with ten problem sizes each. The results prove that the run times increase exponentially with the problem size for standard encoding schemes, while the increase is linear with genetic programming. Genetic programming was up to 722 times faster than the benchmark algorithm. The improvement in computation time does not reduce the algorithm performance in finding optimal solutions; often, it even increases. We conclude that evolving functions enables more efficient land use allocation planning and yields much potential for other spatial optimization applications.

2012 ACM Subject Classification Computing methodologies → Discrete space search

Keywords and phrases Land use planning, Spatial optimization, Solution encoding, Computation time reduction

Digital Object Identifier 10.4230/LIPIcs.GIScience.2023.4

Supplementary Material *Software (Source Code)*: <https://doi.org/10.17632/4tw223jvjv.3>

Dataset (Illustrations): <https://doi.org/10.6084/m9.figshare.21977228.v2>

1 Introduction

Land is scarce, and the competition for land is increasing [4] and continues to increase in the future. Efficient planning can serve social, economic and ecological needs at the same time [4]. In contrast, inefficient and inconsiderate planning has much potential to cause future problems [15]. One aspect of land use planning is the allocation of land use activities.

In order to efficiently allocate land uses, land use planners specify the land use context by defining the land units, the land use categories, the scale, the benefits and undesired outcomes associated with the activities of future land use allocation. Land use modellers can translate these specifications into a solvable model: a land use allocation problem. The

¹ Corresponding author



modeller has to define the decision variable, the constraints, and the objective functions. The decision variable of the optimization is what land use category is assigned to which land unit. Land units can vary in their spatial representation, i.e. vector or raster, and in their encoding scheme.

Currently, the solution scheme in land use allocation optimization is a linear sequence of elements in which every element is one decision variable and is assigned one land use category. One encoding scheme is associating one land use category with one element in the sequence representing one land unit [2, 26]. Another option is to combine multiple neighbouring land units into patches and associate each patch with one element in the sequence [22, 29]. Then, benefits and undesired outcomes are formulated as objective and constraint functions. Constraint functions validate whether a solution violates the defined constraint(s), and objective function(s) quantify the solution's expected benefits.

Optimization algorithms identify solutions to the land use allocation problem. The problem specification determines whether exact algorithms are applicable to solve the problem or whether heuristic approaches are required. If the effort for solving the problem increases exponentially with the number of decision variables, the problem is NP-hard, and heuristic optimization methods are used [27]. Most land use allocation problems fall into the category of NP-hard problems: The number of land units u and the number of land uses categories luc defines the number of possible combinations n of the land use allocation problem: $n = luc^u$. Land use allocation algorithms using the standard encoding are slow when landscapes are complex [28], and face exponentially increasing computation times with increasing problem sizes [27].

Another encoding scheme, yet uncommon in spatial optimization, is a tree that organizes the elements recursively [24]. Since the choice of a suitable encoding has been proven to improve optimization [12] and land use allocation optimization encounters scaling problems with increasing numbers of land units, we propose using the recursive tree encoding. Trees can represent functions, and functions can represent fields [14]: If a function contains two variables, it is possible to represent continuous fields with longitude, latitude, and a variable. Therefore, the tree representation offers an alternative solution encoding scheme to represent spatial objects. Functions describe spatial patterns in the field of geostatistics [23], why should it not be possible to evolve continuous fields as functions to produce favourable land use patterns, for example, patterns that involve spatial compactness, or specific shapes of contiguous land uses?

Much research has been conducted to improve land use allocations with the standard solution encoding, but none on evolving functions to generate land use maps. This study aims to fill the research gap by opening the research domain to using functions as solution encoding. We propose a new method to map functions to nominal land use maps. We compare the new approach with state-of-the-art allocation algorithms on two multi-objective land use allocation problems, one raster and one vector land use allocation problem. In the remainder of this work, we are going to answer the following research questions:

1. How does the computation time of optimizing land use maps represented as functions scale with an increasing number of land units?
2. How does the function-evolving algorithm perform in comparison to state-of-the-art land use allocation algorithms in terms of computation time and the optimal solution quality?

2 Background

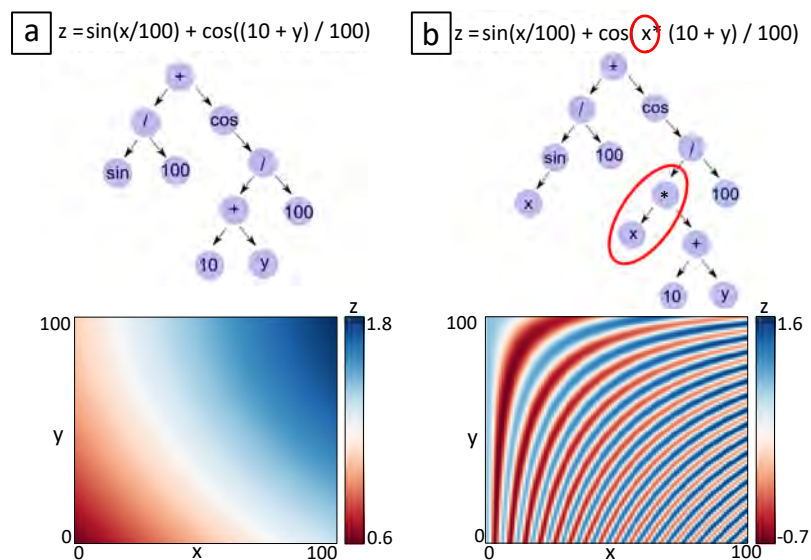
Heuristic search algorithms are most often used to solve land use allocation problems [26]. Heuristic search algorithms identify solutions that are not guaranteed to be truly optimal but help find “good enough solutions” for hard problems in finite time [27]. In contrast to exact optimization algorithms, heuristic optimization algorithms explore the search space of possible solutions until reaching a termination criterion [21].

Common heuristic optimization algorithms for solving land use allocation problems are population-based algorithms, e.g., Genetic Algorithms (GAs) and Particle Swarm Optimization (PSO). GAs and single-objective PSO are commonly applied [26] for single-objective land use allocation problems, and the Non-dominated Sorting Genetic Algorithm II (NSGA 2) [6] is the most often used algorithm for solving multi-objective land use allocation problems [26]. Its successor, the NSGA 3 algorithm, leads to better distributed optimal solutions between conflicting objectives [18]. These algorithms use different search strategies: Genetic Algorithms mimic evolutionary processes by utilizing fitness proportionate selection and genetic recombinations of individuals within a population [10]. In PSO, the equivalent of an individual in a population is a particle in a swarm of particles, moving within the problem space [17] to find the best positions.

The algorithms evolve solutions by manipulating the solution. In land use allocation algorithms, the manipulation procedures of the algorithms are either applied on the sequences containing single land units [1] or of land use patches [16]. One advantage of using patches in comparison to single land units is the lower number of decision variables [29]. Another advantage of evolving patches is the higher likelihood of obtaining solutions with innate spatial relationships like adjacency or connectivity [17], which are often desired characteristics in land use allocation [26]. Numerous operators have been developed to steer the optimization process towards patches with certain characteristics such as compactness [17], or validity [29]. It is important to notice that some manipulations are computationally more demanding than others, but all manipulations of solutions with the common land use map encoding lead to an increased computational effort when considering more land units.

On the other hand, genetic programming (GP) is an evolutionary algorithm that evolves solutions with a different encoding: program trees that build functions [13]. The encoding yields the potential to evolve solutions detached from single land units by evolving fields that represent the whole landscape as one object: If the functions incorporate spatial variables, e.g. the latitude and longitude, the function produces an output variable for any given position. Other components of the function can influence the output variable. Combined, the spatial and non-spatial components define how the variable varies in space. Therefore, it is possible to optimize the spatial variation of the output variable by manipulating the non-spatial components. Since the output variable is detached from land units, the number of land units does not affect the computational effort when manipulating the solutions.

The algorithm has been applied to a wide variety of non-spatial problems [20], but neither to land use allocation problems nor to spatial optimization problems in general. GP yields better results than GA in related applications, e.g. for generating grids of a continuous variable for photomosaics [19]. One identified reason is the higher flexibility due to the encoding of solutions, where little adaptations of the program trees can lead to many changes in the produced grid and potentially towards favourable patterns [19]. In addition to producing optimal grids of a continuous variable, genetic programming also proved to perform well on discrete variable classification [11]. The promising results of these studies suggest that genetic programming is applicable to allocating land use.



■ **Figure 1** Two exemplary individuals with the function, the program tree, and the resulting field. The primitives are sine, cosine, addition, subtraction, multiplication, and division. The terminals are 10 and 100, and the x and y inputs range from 1 to 100, resulting in a continuous z value.

3 Methods

3.1 Land use allocation optimization using genetic programming

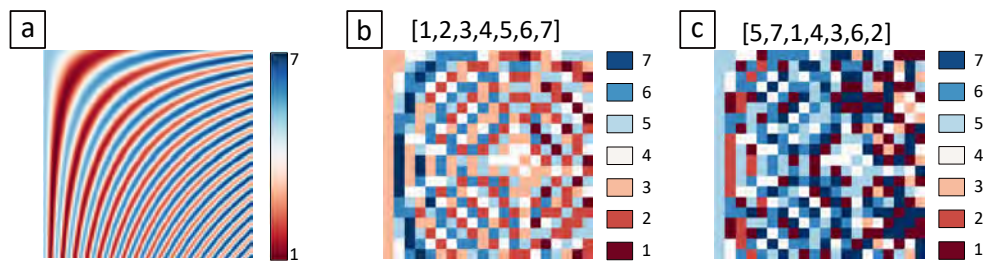
Generating fields with genetic programming

In genetic programming, every individual of the population is a “hierarchical composition of primitive functions and terminals” [13]. Typically, arithmetic operations, mathematical functions, or conditional logical operations constitute the functions [13]. The terminals and numeric constants are inputs to the problem. In our case, where solutions to the problems are two-dimensional fields, the inputs are x and y coordinates. The coordinates are two input variables that can repeatedly appear in the program trees (Fig. 1). When incorporating the coordinates within into mathematical functions, spatial For illustration purposes, the individuals are visualized as program trees (Fig. 1).

Mapping continuous fields to nominal land use maps

First, we retrieve the input coordinates from the land units. In the case of a raster representation of land uses, the row and column IDs serve as the x and y inputs of the program trees. In the case of a vector land use representation, the centroid coordinates of the land units serve as x and y inputs. Applying the function on the x and y inputs of the program trees defines the output variable z (Fig. 1). We use the mean of z per patch for the patch representation. Then z is min-max normalized to a range that matches the land use categories. Finally, rounding the normalized z-values to integers generates the desired nominal values.

This mapping procedure suffices to retrieve nominal values per land unit. However, the continuous variable contains an order, and mapping the continuous variable to a nominal variable propagates an order. It is not particularly meaningful to define an order between land uses urban, forest, or pasture. If this order were ignored, then the likelihood of neighbouring land uses would be influenced by the predefined land use order. To avoid this artifact, we



■ **Figure 2** A continuous field (a), mapped to two nominal maps (b,c) with different land use category orders.

actively handle the orders of land use orders categories in the optimization. Every individual gets assigned a land use order element that contains randomly shuffled land use category IDs (Fig. 2, b and c). The obtained integer values are re-mapped with each individual's land use order (Fig. 2). With this approach, the same function (Fig. 1, b) results in the same continuous field (Fig. 2, a), but the nominal values differ. Without re-mapping, land use with id 1 would always have a higher likelihood of neighbouring to land use 2 than to land use 5. An association of different land use orders to individuals within the population leaves the potential to find an optimal combination of land use orders and functions in the optimization.

GP procedure for the multi-objective land use allocation

The algorithm procedure starts with a random initialization of individuals until reaching the population size. This study uses the standard initialization called *ramped half and half*. It is a combination of two tree-generation algorithms *grow* and *full*, and in both the primitives and terminals are generated at random [13]. The *grow* initialization creates a sub-tree with a tree depth that is also randomly selected between a minimum and a maximum tree depth threshold. In contrast, the *full* algorithm generates a sub-tree with a depth that equals a depth threshold. Then, until a termination criterion is reached, in every generation, the algorithm evaluates the individuals with the objective function(s) and constraint(s), selects individuals for reproduction with a selection operation, generates offspring individuals in a crossover operation, and mutates the individuals in a mutation operation.

Since the algorithm is applied to multi-objective land use allocation problems, the selection operation selects individuals based on multiple objective values. We use the selection procedure from the NSGA 3 algorithm [7] that is based on the principle of Pareto efficiency and is designed to find individuals close to desired reference points. Reference points can be user-defined or distributed strategically, for example, using equal distances on the hyper-plane [5]. We refer to the original paper for a detailed description for details [7].

We use the standard GP operators *one-point crossover* and *one-point mutations* for the crossover and mutation. For example, in the *one-point crossover*, a common crossover point in the parent solutions is selected randomly, and then the corresponding sub-trees are exchanged [25]. In the *one-point mutation*, a random point of the tree is selected and then replaced with a newly generated sub-tree.

3.2 Land use allocation test problems

We use two land use allocation problems (Tab. 1) for testing the proposed method. The first test problem is a synthetic raster land use problem with 8 land use categories, two constraints, and four maximization objectives. Both problems are multi-class combinatorial

■ **Table 1** Land use problem specifications. The raster problem is re-used; for more details see [29]. The vector problem is designed for this study.

Raster problem	Vector problem
Data and spatial representation	
Synthetic raster data serves as an initial land use map. The problem can be approached with single raster cell representation and raster patches.	Real-world parcels (vector) serve as spatial units and for the initial land use map. Sixteen land use categories associated with the parcels are mapped into seven land uses.
Land use categories	
Cropland 1-5, representing five different levels of agricultural productivity, pasture, forest, urban	Civil, rural non-forest, industrial, agriculture, forest, residential, and the last combines transport and water.
Constraints	
<i>Land use transition constraint</i> The transition of urban land use is restricted, forests can only be converted to pasture, and pasture cannot be converted.	<i>Land use transition constraint</i> Transitions of civil, and water and transport land uses are restricted; only rural-non forest be converted to forest.
<i>Area proportion constraint</i> Permitted ranges of 10-25% for forest, 10-30% for pasture. No area proportion constraint for other land uses.	<i>Area proportion constraint</i> Permitted ranges of 0-50% for industrial, 10-80% for agriculture, 15-100% for forest, 10-100% for residential. No area proportion constraint for other land uses.
Objective functions	
<i>Max. species richness (SR)</i> An empiric value that changes with the total forest area (unitless)	<i>Max. urban compactness (UC)</i> Count of adjacencies between land units of the categories civil, residential, and industrial.
<i>Max. habitat heterogeneity (HH)</i> Sum over edges between different land use types, where low-intensity land uses get higher weights than high-intensity land uses (unitless)	<i>Max. agriculture within water range (AW)</i> Area in <i>ha</i> of land use agriculture that intersects with 500-meter buffers around waters.
<i>Max. water yield (WY)</i> Relative differences in evapotranspiration rates between land use types (unitless)	<i>Max. Contiguous agriculture size (AS)</i> Average patch size in <i>ha</i> of contiguous agriculture.
<i>Max. crop yield (CY)</i> Sum of all logarithmic products of cropland intensity and soil fertility over all cells (unitless).	<i>Max. distance residential to wind plants (DRW)</i> Average distance of residential areas to the closest wind plant point in <i>km</i> .

problems; the decision variables are elements in a sequence with a length that equals the number of land units. Each element is associated with one land use category, represented as an integer value. For more specifications about the problem background and formulation, we refer to Tab. 1 and [29]. Both problems have initial land use maps. The problem instance classes and the initial land use maps are available online ².

The second test problem is a vector land use problem with 7 land use categories, two constraints and four maximization objectives [30]. We constructed the problem for testing the algorithm's performance with parcels located in Germany.

² <https://data.mendeley.com/datasets/4tw223jvfv>

Furthermore, we generate the single objective optimal land use configurations per objective. For example, we allocated only land use *Cropland 5* while not violating the constraints to generate the optimal solution for the objective *Crop Yield*, and only the land uses *Civil*, *Industrial* and *Residential* for the objective *Urban Compactness*. The only exception is the objective *Habitat Heterogeneity*, for which we approximate the single objective optimal solution. These single objective optima are the extreme ends of the Pareto fronts. Therefore, they are insufficient to determine whether an algorithm finds the true Pareto front. However, it serves as an indicator to determine whether or not an algorithm can find optimal solutions or how far it is off from the known optima.

3.3 Design of simulation experiments and software availability

■ **Table 2** Simulation experiment with a) Run time analysis over 10 problem sizes. b) Single-objective best solutions found by the algorithms and the known optima.

a) Run time analysis				
Problem type	Problem size	Algorithm	Nr. of generations	Pop. size
Raster	100 - 22500	GP	10	40
Raster	100 - 22500	NSGA 2 with repair mutation	10	40
Raster	100 - 22500	NSGA 2 no repair mutation	10	40
Vector	2075 - 13687	GP	10	40
Vector	2075 - 13687	NSGA 3	10	40
b) Single-objective solution comparison				
Problem type	Problem size	Algorithm	Nr. of generations	Pop. size
Raster	100	GP	100	200
Raster	100	NSGA 2 with repair mutation	100	200
Raster	10,000	GP	100	200
Raster	1,000,000	GP	100	200
Vector	13687	GP	100	200
Vector	13687	NSGA3	100	200

In the first experiment, we perform a benchmark between GP and the most commonly multi-objective land use allocation algorithm NSGA 2 on the multi-objective raster land use problem (Tab. 3). The NSGA 2 can not be applied without adaptations for solving land use allocation problems. Therefore, we compare GP to a land use allocation algorithm that bases on NSGA 2 on the multi-objective raster land use problem defined in [29]. The authors suggest a multi-objective land use allocation algorithm (CoMOLA) to solve the land use problem. The algorithm offers the option to use a repair mutation operation for patches. The spatially explicit repair functions can improve the search for optimal solutions by repairing infeasible individuals. We perform a run time benchmark on 10 problem sizes with raster dimensions from 10*10 to 150*150 cells with a step size of 10. Then, to indicate the algorithm performance of finding optimal solutions, we compare the best solutions of the single objectives from both algorithms on the 10*10 problem size to the known single objective optima.

In the second simulation experiment, we test the algorithm performance on the multi-objective vector land use problem with features representing land units. We select the NSGA 3 algorithm, the successor of NSGA 2, as benchmark algorithm for two reasons. First, we use the same selection procedure [7]. Second, NSGA 3 has proven its ability to find better-distributed solutions in Pareto fronts and has been successfully applied to solving land use allocation problems [18]. We perform a run time benchmark on 10 problem sizes ranging from 2075 to 13687 land units and compare the single objective optimal solutions for 2075 and for 13687 land units to the known single objective optima. The software used is open source and the results are fully reproducible.

The code, input data, and results files are available at Mendeley Data².

4 Results

4.1 Raster land use allocation problem

The scaling potential of the run times is promising. While CoMOLA with the patch repair mutation took 171 minutes to evaluate 200 individuals in 100 generations, GP needed 5 seconds for the same number of evaluations. The larger the problem instances, the larger the difference between the run times. While the run times of the CoMOLA based on the NSGA 2 algorithm increase exponentially with increasing raster problem sizes, GP run times increase linearly (Fig. 3, a).

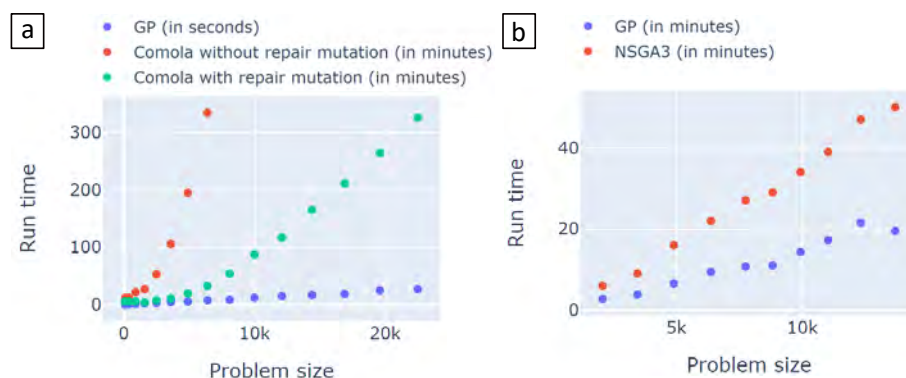


Figure 3 Total run times of with increasing land use problem sizes. a) Raster land use problem ranging from 100 (10x10) to 22500 (150x150). b) Vector land use problem with problem sizes from 2075 to 13687 land units.

The highest difference, therefore, was observed on the largest raster problem instance with 22500 (150 * 150) grid cells: here, NSGA 2 required 325 minutes, whereas GP required 27 seconds, which is 722 times faster. When applying the repair mutation on CoMOLA, the difference is even higher. The spatially explicit repair function lead to computation times that exceeded 5 hours at a problem size with 80*80 cells. In comparison: When testing GP on the problem with 1000*1000 cells leading to one million decision variables, the algorithm took 742 minutes.

The single-objective optimal solutions derived with GP (Fig. 4 and Tab. 3) prove that the algorithm can and does find global optima and solutions close to the global optimum. For obtaining the optimal solution for objective *Crop yield*, only one pixel (Fig. 4 a, top left corner) is off, where cropland 4 is allocated instead of cropland 5. All other non-constrained land uses are set to the optimal land use cropland 5. The same applies to the finer resolution of 100 * 100 pixels, where 15 out of 10000 pixels are not set to the optimal land use (Fig. 4b,

■ **Table 3** Single objective extreme values obtained with the algorithms with the known global optima.

Raster	Size	CY [-]	HH [-]	SR [-]	WY [-]
NSGA 2 with repair mutation	10*10	125.7	84.2	9.51	97.7
GP	10*10	134.7	282.2	9.51	98.0
Known optimum	10*10	138.2	354	9.51	98.9
NSGA 2 with repair mutation	100*100	-	-	-	-
GP	100*100	13,574	24,903	23.9	9,890
Known optimum	100*100	13,615	34,778	23.9	9891
NSGA 2 with repair mutation	1000*1000	-	-	-	-
GP	1,000*1,000	1,359,403	2,703,280	60.0	989,107
Known optimum	1,000*1,000	1,359,404	3,471,380	60.0	989,108
Vector	Size	UC [-]	AS [ha]	DRW [km]	AW [ha]
NSGA3	2075	313	755	0.021	725
GP	2075	347	738	0.023	770
Known optimum	2075	372	1144	0.027	982
NSGA3	13,687	1,659	7,074	0.894	2,787
GP	13,687	1907	6,257	1.135	2,731
Known optimum	13,687	2,211	9,321	1.23	3,830

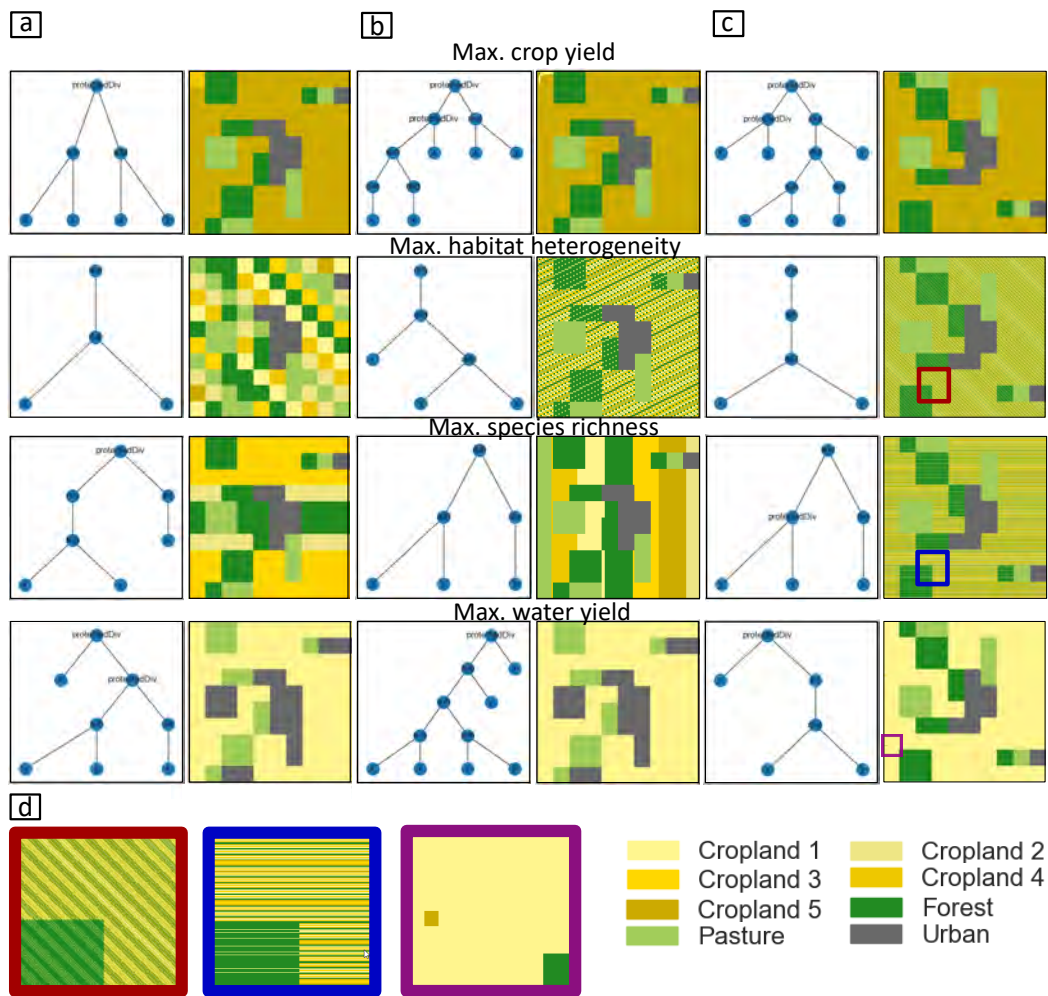
top left corner). The global optimum was obtained on both spatial resolutions for objective *Water yield* with Cropland 1 being the best land use. For objective *Species Richness*, the global optimum is obtained, but this is comparatively easy to obtain by reaching 25% of land use forest since it corresponds to the upper area constraint for land use forest. More remarkable is the produced cluster in optimal solutions for objective *Habitat heterogeneity*. For this objective, the perfect land use pattern is produced when the number of neighbours between the constrained land use forest, pasture and cropland use 1 is maximized, followed by neighbours to cropland 2 etc. GP found this pattern (Tab. 3) that seems impossible to find by CoMOLA: The best objective value obtained with GP is 3.35 times higher than the best objective value obtained by CoMOLA.

Moreover, GP is not negatively affected by larger problem instances; the convergence to the single objective optima is even better on the larger problem with 100*100 cells compared to the small problem with 10*10 cells. Even on the largest problem with 1000*1000 cells, GP found one single objective global optima and two solutions that deviate 0.001% and 0.00001% from the global optima (Tab. 3). This observation indicates the scaling potential of the algorithm's performance on a finer spatial resolution.

The single objective optima show that GP can find optimal spatial patterns for objective functions based on adjacency and connectivity for small and large land use allocation problem instances.

4.2 Vector land use allocation problem

The optimization of the vector problem requires more computation time compared to the raster problem. The computationally more expensive fitness evaluations, in which intersections, distance, and adjacency operations on features are used, are the reason for the longer run times. However, the field-evolving GP is considerably faster than the NSGA 3, and the difference increases with more land units (Fig. 3, b). On average, GP is 138% faster regardless



■ **Figure 4** Single objective optimal solutions for raster land use problems with problem sizes 10x10 cells (a), 100x100 cells (b), and 1000*1000 cells (c). Close-ups (d) show produced patterns from selected regions of the 1000*1000 cell maps: The red frame shows the close-up for objective Max. habitat heterogeneity, the blue frame shows the close-up for objective Max. species richness and the purple frame shows the close-up for the objective Max. water yield.

of the problem size. However, the run time of the GP also scales well on larger problems. On the larger problem size with 13687 land units, GP took, on average, 61% less computation time per land unit than on the problem with 2075 land units.

GP did not find the global optima for the objectives in the vector problem (Tab. 3). However, GP also outperforms the NSGA 3 algorithm on the land use problem with 2075 land units (Tab. 3). The single objective optimal values of *Urban Compactness* (UC), and *Agriculture in water range* (AW) are 9.2%, 9.5%, and 6.2% better. NSGA 3 found a 2.3% better single objective optima for the *Contiguous agriculture size*. The number of optimal solutions is also higher, with 57 compared to just 7 obtained with the NSGA 3. Furthermore, GP found solutions (Fig. 5) that show spatial patterns, such as contiguous agriculture land uses, or the seemingly ordered land uses along horizontal (Fig. 5 b, last row) and the vertical axis (Fig. 5 a, last row, and b, third row)³.

³ Additional results, including Pareto frontiers, are available with a DOI at figshare.

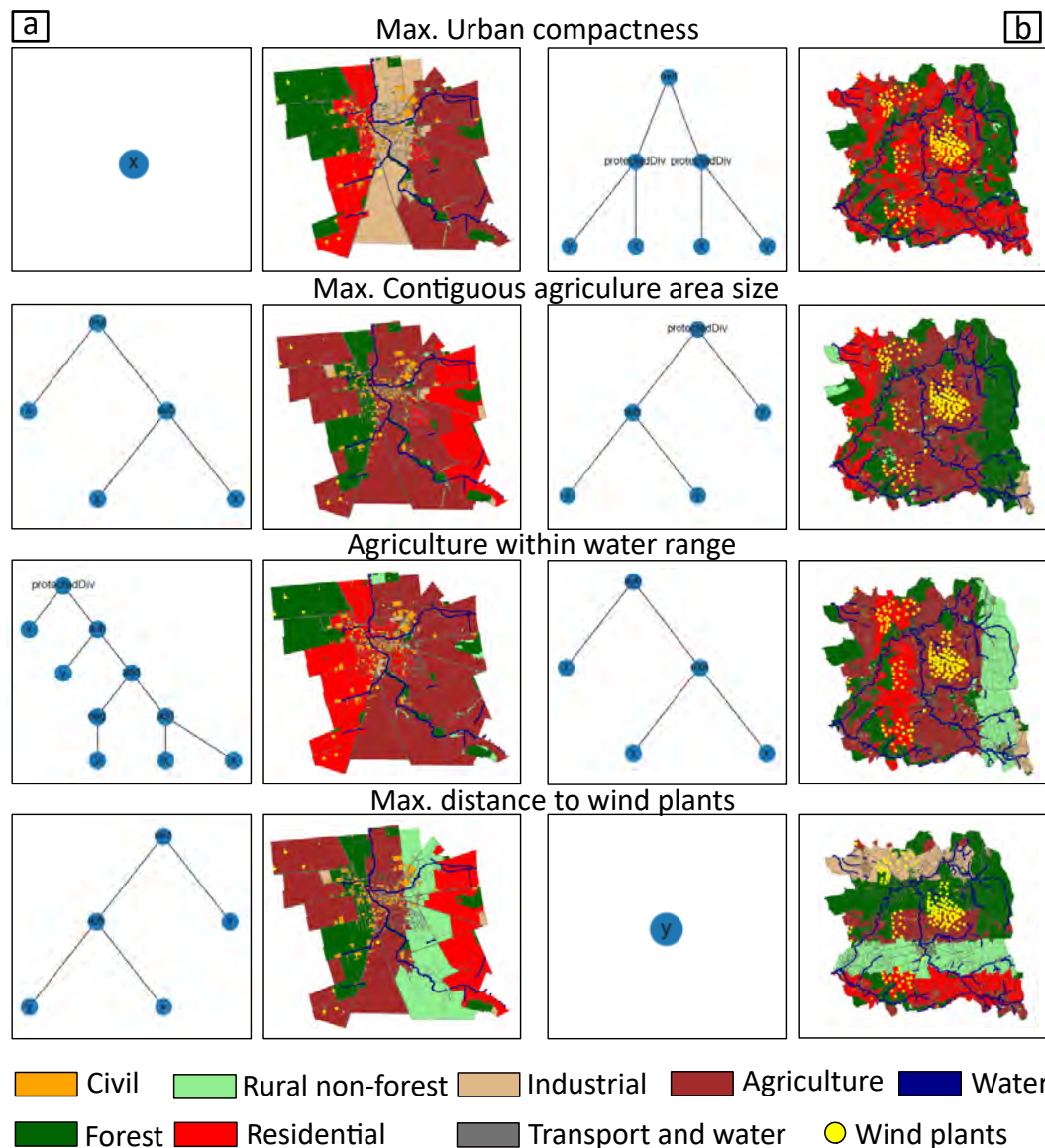


Figure 5 Single objective optimal solutions for vector land use problems with two problem sizes 2075 parcels (run time: 246 minutes) and 13687 parcels (run time: 761 minutes).

5 Discussion

5.1 Potential of encoding spatial objects as functions

The results of this study show that optimizing functions that generate continuous fields can lead to more optimal land use configurations in shorter computation times compared to using algorithms with standard encoding. The optimal land use maps produced with functions in the GP algorithms are closer to the global optima, and in many cases, GP even found the global optima. The observed scaling shows the potential for high-resolution land use units and/or larger study areas, which is promising for other land use allocation problems than the ones shown here. Another example is uncertainty analysis of land use allocation optimizations, which require many optimization executions and benefit even more from the

decreased computational cost [8]. Other spatial optimization problems might also be solved with the GP algorithm, e.g. 3D routing optimizations for which a sequence of 3D points is optimized instead of evolving functions [9], or facility location planning [3]. Possibly, GP can be applied to solve spatial problems that change over time by including a time dimension variable as part of the functions.

5.2 Limitations and future work

In this work, we used functions that include spatial dependencies in both x and y directions in the encoding of solutions, while a sequence of elements that represent spatial units does not. This yields a great advantage for spatial optimization problems that handle spatial objects and offers much potential for future investigation. This approach comes with disadvantages, too, e.g. the necessity to attach a random land use order to every solution to mitigate the effect of translating a continuous to a nominal variable. Investigating the random land use order association with individuals in more detail is, therefore, important for future research. For example, in our results, the portion of unique land use orders decreases over the generation and stabilizes at 40 after 50 generations. Finding out whether the observed behaviour is an anomaly or whether some land use orders are particularly suitable for solving the problem may yield important insights.

In this study, we used standard GP initialization, crossover and mutation operators and no hyper-parameter tuning to prove the general applicability of the GP algorithm on land use allocation problems. Many different initializations of trees, mutations, or crossover exist for which many parameter settings are possible, and some operators and parameter settings may yield better results for land use allocation problems or other spatial optimization problems. One parameter that should be tuned is the maximum tree depth. This parameter was set to 8, but the maximum tree depth in the optimal solutions was 5. In the vector problem, the tree depth was even shallower; some only had one terminal (Fig. 5). The parameter tuning is, therefore, future work to further improve the algorithm performance.

Lastly, the better performance on the raster problem compared to the vector problem leaves room for further analysis. The static boundaries of the features might be the reason for this observation: While GP could generate patterns and clusters that potentially benefit objectives in the raster case, that positive characteristic of the algorithm can not be realized in the vector case where the object extents are set. Another reason may be the usage of polygon centroids as x and y inputs to the function. A different mapping is possibly better for considering the whole feature's extent, e.g., using multiple points per polygon as input.

6 Conclusion

Standard land use allocation optimization algorithms cannot cope with large land use allocation problems due to the solution encoding. Using function as solution encoding proved to solve land use problems more efficiently. The functions represent spatial fields that are mapped to nominal land use maps. We solve the identified mapping problem from continuous fields to nominal maps by associating random land use orders with the individuals of the GP population.

GP proved its ability to alleviate exponentially increasing run times of the standard encoding scheme on a raster and a vector problem. While the computation time using the standard solution encoding increased exponentially, the computation time using GP increased linearly. As a consequence, the reduction of computation time increases exponentially with larger problem instances, too. On the largest raster problem instance, GP was up to 722

times faster than the NSGA 2 land use allocation algorithm. The difference in computation time further increases when comparing GP to the standard encoding coupled with spatially explicit operators.

Moreover, the improvement in computation time does not affect the algorithm's performance in finding better solutions than the benchmark algorithms. GP obtained better single-objective solutions than NSGA 2 and NSGA 3 on six out of eight objectives of the two benchmark problems. Moreover, GP found the global single objective optima for three objectives of the raster problem with 10*10 cells and 100*100 cells. Even on the 1000*1000 single-cell raster problem, one global optimum was found and two near-optimal (deviation of 0.001% and 0.00001% from global optima). The highest increased performance was obtained for the objective *Habitat Heterogeneity* of the raster problem that requires finding a highly complex spatial pattern of adjacent land uses. Also, GP found contiguous clusters required to find optimal solutions to four other objectives. This shows that GP can produce land use maps with spatial patterns that involve adjacency and connectivity.

We conclude that evolving functions enable more efficient land use allocation optimizations in the future and that the approach is a promising method for other spatial optimization problems.

References

- 1 Kai Cao, Bo Huang, Shaowen Wang, and Hui Lin. Sustainable land use optimization using boundary-based fast genetic algorithm. *Computers, Environment and Urban Systems*, 36(3):257–269, 2012. doi:10.1016/j.compenvurbsys.2011.08.001.
- 2 Kai Cao, Muyang Liu, Shu Wang, Mengqi Liu, Wenting Zhang, Qiang Meng, and Bo Huang. Spatial multi-objective land use optimization toward livability based on boundary-based genetic algorithm: A case study in singapore. *ISPRS International Journal of Geo-Information*, 9(1):40, 2020. doi:10.3390/ijgi9010040.
- 3 Richard L. Church and Alan T. Murray. *Business Site Selection, Location Analysis and GIS*. John Wiley & Sons, Inc, Hoboken, NJ, USA, 2008. doi:10.1002/9780470432761.
- 4 Felix Creutzig, Christopher Bren d'Amour, Ulf Weddige, Sabine Fuss, Tim Beringer, Anne Gläser, Matthias Kalkuhl, Jan Christoph Steckel, Alexander Radebach, and Ottmar Edenhofer. Assessing human and environmental pressures of global land-use change 2000–2010. *Global Sustainability*, 2, 2019. doi:10.1017/sus.2018.15.
- 5 Indraneel Das and John E. Dennis. Normal-boundary intersection: A new method for generating the pareto surface in nonlinear multicriteria optimization problems. *SIAM Journal on Optimization*, 8(3):631–657, 1998. doi:10.1137/S1052623496307510.
- 6 Kalyanmoay Deb, Amrit Pratap, Sameer Agarwal, and T. Meyarivan. A fast and elitist multiobjective genetic algorithm: Nsga-ii. *IEEE Transactions on Evolutionary Computation*, 6(2):182–197, 2002. doi:10.1109/4235.996017.
- 7 Kalyanmoy Deb and Himanshu Jain. An evolutionary many-objective optimization algorithm using reference-point-based nondominated sorting approach, part i: Solving problems with box constraints. *IEEE Transactions on Evolutionary Computation*, 18(4):577–601, 2014. doi:10.1109/TEVC.2013.2281535.
- 8 Moritz Hildemann and Judith A. Versteegen. Quantifying uncertainty in pareto fronts arising from spatial data. *Environmental Modelling & Software*, 141:105069, 2021. doi:10.1016/j.envsoft.2021.105069.
- 9 Moritz Hildemann and Judith A. Versteegen. 3d-flight route optimization for air-taxis in urban areas with evolutionary algorithms and gis. *Journal of Air Transport Management*, 107:102356, 2023. doi:10.1016/j.jairtraman.2022.102356.
- 10 John H. Holland. *Adaptation in natural and artificial systems: An introductory analysis with applications to biology, control, and artificial intelligence*. u Michigan Press, 1975.

- 11 Vijay Ingalalli, Sara Silva, Mauro Castelli, and Leonardo Vanneschi. A multi-dimensional genetic programming approach for multi-class classification problems. In Miguel Nicolau, Krzysztof Krawiec, Malcolm I. Heywood, Mauro Castelli, Pablo García-Sánchez, Juan J. Merelo, Victor M. Rivas Santos, and Kevin Sim, editors, *Genetic Programming*, volume 8599 of *Lecture Notes in Computer Science*, pages 48–60. Springer Berlin Heidelberg, Berlin, Heidelberg, 2014. doi:10.1007/978-3-662-44303-3_5.
- 12 Konstantin Klemm, Anita Mehta, and Peter F. Stadler. Landscape encodings enhance optimization. *PloS one*, 7(4):e34780, 2012. doi:10.1371/journal.pone.0034780.
- 13 John R. Koza. Genetic programming as a means for programming computers by natural selection. *Statistics and Computing*, 4(2), 1994. doi:10.1007/BF00175355.
- 14 Werner Kuhn. Core concepts of spatial information for transdisciplinary research. *International Journal of Geographical Information Science*, 26(12):2267–2276, 2012. doi:10.1080/13658816.2012.722637.
- 15 Arika Ligmann-Zielinska, Richard L. Church, and Piotr Jankowski. Spatial optimization as a generative technique for sustainable multiobjective land-use allocation. *International Journal of Geographical Information Science*, 22(6):601–622, 2008. doi:10.1080/13658810701587495.
- 16 Hongjiang Liu, Fengying Yan, and Hua Tian. Towards low-carbon cities: Patch-based multi-objective optimization of land use allocation using an improved non-dominated sorting genetic algorithm-ii. *Ecological Indicators*, 134:108455, 2022. doi:10.1016/j.ecolind.2021.108455.
- 17 Yaolin Liu, Jinjin Peng, Limin Jiao, and Yanfang Liu. Psola: A heuristic land-use allocation model using patch-level operations and knowledge-informed rules. *PloS one*, 11(6):e0157728, 2016. doi:10.1371/journal.pone.0157728.
- 18 Jamshid Maleki, Zohreh Masoumi, Farshad Hakimpour, and Carlos A. Coello Coello. Many-objective land use planning using a hypercube-based nsga-iii algorithm. *Transactions in GIS*, 26(2):609–644, 2022. doi:10.1111/tgis.12876.
- 19 Shahrul Badariah Mat Sah, Vic Ciesielski, Daryl D’Souza, and Marsha Berry. Comparison between genetic algorithm and genetic programming performance for photomosaic generation. In Xiaodong Li, Michael Kirley, Mengjie Zhang, David Green, Vic Ciesielski, Hussein Abbass, Zbigniew Michalewicz, Tim Hendtlass, Kalyanmoy Deb, Kay Chen Tan, Jürgen Branke, and Yuhui Shi, editors, *Simulated Evolution and Learning*, volume 5361 of *Lecture Notes in Computer Science*, pages 259–268. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008. doi:10.1007/978-3-540-89694-4_27.
- 20 James McDermott, David R. White, Sean Luke, Luca Manzoni, Mauro Castelli, Leonardo Vanneschi, Wojciech Jaskowski, Krzysztof Krawiec, Robin Harper, Kenneth de Jong, and Una-May O’Reilly. Genetic programming needs better benchmarks. In Terence Soule and Jason H. Moore, editors, *Proceedings of the 14th annual conference on Genetic and evolutionary computation*, pages 791–798, New York, NY, USA, 07072012. ACM. doi:10.1145/2330163.2330273.
- 21 Ibrahim H. Osman and James P. Kelly, editors. *Meta-Heuristics*. Springer US, Boston, MA, 1996. doi:10.1007/978-1-4613-1361-8.
- 22 Tingting Pan, Yu Zhang, Fenzhen Su, Vincent Lyne, Fei Cheng, and Han Xiao. Practical efficient regional land-use planning using constrained multi-objective genetic algorithm optimization. *ISPRS International Journal of Geo-Information*, 10(2):100, 2021. doi:10.3390/ijgi10020100.
- 23 Edzer J. Pebesma. The role of external variables and gis databases in geostatistical analysis. *Transactions in GIS*, 10(4):615–632, 2006. doi:10.1111/j.1467-9671.2006.01015.x.
- 24 Fernando Peres and Mauro Castelli. Combinatorial optimization problems and metaheuristics: Review, challenges, design, and development. *Applied Sciences*, 11(14):6449, 2021. doi:10.3390/app11146449.
- 25 Riccardo Poli and W. B. Langdon. Genetic programming with one-point crossover. In P. K. Chawdhry, R. Roy, and R. K. Pant, editors, *Soft Computing in Engineering Design and Manufacturing*, pages 180–189. Springer London, London, 1998. doi:10.1007/978-1-4471-0427-8_20.

- 26 Mostafizur Rahman and György Szabó. Multi-objective urban land use optimization using spatial data: A systematic review. *Sustainable Cities and Society*, 74:103214, 2021. doi:10.1016/j.scs.2021.103214.
- 27 Franz Rothlauf. Optimization methods. In Franz Rothlauf, editor, *Design of Modern Heuristics*, Natural Computing Series, pages 45–102. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011. doi:10.1007/978-3-540-72962-4_3.
- 28 Mingjie Song and DongMei Chen. A comparison of three heuristic optimization algorithms for solving the multi-objective land allocation (mola) problem. *Annals of GIS*, 24(1):19–31, 2018. doi:10.1080/19475683.2018.1424736.
- 29 Michael Strauch, Anna F. Cord, Carola Pätzold, Sven Lautenbach, Andrea Kaim, Christian Schweitzer, Ralf Seppelt, and Martin Volk. Constraints in multi-objective optimization of land use allocation – repair or penalize? *Environmental Modelling & Software*, 118:241–251, 2019. doi:10.1016/j.envsoft.2019.05.003.
- 30 Daoqin Tong and Alan T. Murray. Spatial optimization in geography. *Annals of the Association of American Geographers*, 102(6):1290–1309, 2012. doi:10.1080/00045608.2012.685044.

Visualizing Geophylogenies – Internal and External Labeling with Phylogenetic Tree Constraints

Jonathan Klawitter  



University of Auckland, New Zealand

Felix Klesen 

Universität Würzburg, Germany

Joris Y. Scholl

Ruhr-Universität Bochum, Germany

Thomas C. van Dijk  

Ruhr-Universität Bochum, Germany

Alexander Zaft

Universität Würzburg, Germany

Abstract

A *geophylogeny* is a phylogenetic tree where each leaf (biological taxon) has an associated geographic location (site). To clearly visualize a geophylogeny, the tree is typically represented as a crossing-free drawing next to a map. The correspondence between the taxa and the sites is either shown with matching labels on the map (internal labeling) or with *leaders* that connect each site to the corresponding leaf of the tree (external labeling). In both cases, a good order of the leaves is paramount for understanding the association between sites and taxa. We define several quality measures for internal labeling and give an efficient algorithm for optimizing them. In contrast, minimizing the number of leader crossings in an external labeling is NP-hard. We show nonetheless that optimal solutions can be found in a matter of seconds on realistic instances using integer linear programming. Finally, we provide several efficient heuristic algorithms and experimentally show them to be near optimal on real-world and synthetic instances.

2012 ACM Subject Classification Human-centered computing → Geographic visualization; Applied computing → Biological networks; Theory of computation → Discrete optimization

Keywords and phrases geophylogeny, boundary labeling, external labeling, algorithms

Digital Object Identifier 10.4230/LIPIcs.GIScience.2023.5

Related Version *Full Version*: <https://arxiv.org/abs/2306.17348> [16]

Supplementary Material *Software (Source Code)*: <https://www.github.com/joklawitter/geophylo>

Funding *Jonathan Klawitter*: Beyond Prediction Data Science Research Programme (MBIE grant UOAX1932).

Thomas C. van Dijk: DFG grant Di2161/2-1.

1 Introduction

A *phylogeny* describes the evolutionary history and relationships of a set of taxa such as species, populations, or individual organisms [25]. It is one of the main tasks in phylogenetics to infer a phylogeny for some given data and a particular model. Most often, a phylogeny is modelled and visualized with a *rooted binary phylogenetic tree* T , that is, a rooted binary tree T where the leaves are bijectively labeled with a set of n taxa. For example, the phylogenetic tree in Figure 1a shows the evolutionary species tree of the five present-day kiwi (*Apteryx*) species. The tree is conventionally drawn with all edges directed downwards to the leaves and



© Jonathan Klawitter, Felix Klesen, Joris Y. Scholl, Thomas C. van Dijk, and Alexander Zaft; licensed under Creative Commons License CC-BY 4.0

12th International Conference on Geographic Information Science (GIScience 2023).

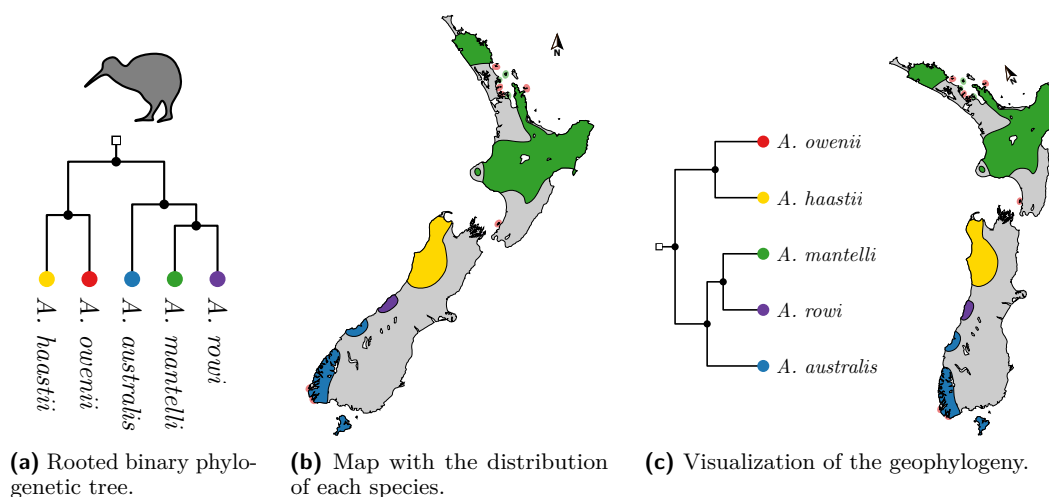
Editors: Roger Beecham, Jed A. Long, Dianna Smith, Qunshan Zhao, and Sarah Wise; Article No. 5; pp. 5:1–5:16

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

5:2 Visualizing Geophylogenies



■ **Figure 1** To visualize Weir et al.’s geophylogeny of the five present-day kiwi species [27], we combine the phylogenetic tree (a) with the distribution map (b) into a single figure (c).

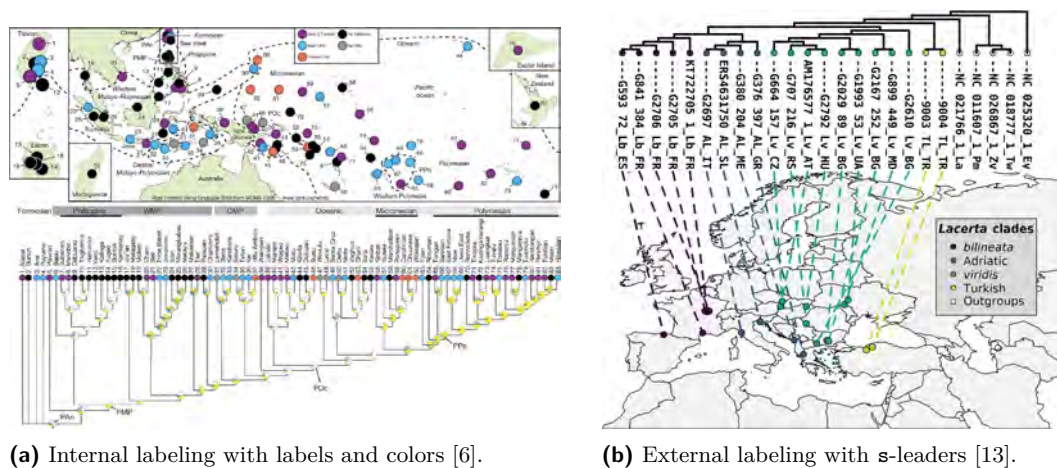
without crossings (*downward planar*). There exist several other models for phylogenies such as the more general phylogenetic networks and unrooted phylogenetic trees; here we only consider rooted binary phylogenetic trees and refer to them simply as phylogenetic trees.

In the field of phylogeography, geographic data is used in addition to genetic data. We may thus have spatial data associated with each taxon such as the distribution range of each species or the sampling site of each voucher specimen used in a phylogenetic analysis. For example, Figure 1b shows the distributions of the kiwi species from Figure 1a. We speak of a *geophylogeny* (or *phylogeographic tree*) if we have a phylogenetic tree T , a map region R , and a set P of features on R that correspond one-to-one with the taxa in T ; see Figure 1c for a geophylogeny of the kiwi species. In this paper, we focus on the case where P is a set of points, called *sites*.

Visualizing Geophylogenies

When visualizing a geophylogeny, we may want to display its tree and its map together in order to show the connections (or the non-connections) between the leaves and the sites. For example, we may want to show that the taxa of a certain subtree are confined to a particular region of the map or that they are widely scattered. In the literature, we mainly find three types of drawings of geophylogenies. In a *side-by-side* drawing, the tree is drawn planar directly next to the map. To show the correspondences between the taxa and their sites, the sites are either labeled or color coded (as in Figure 2a and Figure 1c, respectively), or the sites are connected with *leaders* to the leaves of the tree (as in Figure 2b). We call this *internal labeling* and *external labeling*, respectively. There also exist *overlay* illustrations where the phylogenetic tree is drawn onto the map in 2D or 3D with the leaves positioned at the sites [15, 29], but for brevity we omit further discussion of this style.

Drawing a geophylogeny involves various subtasks, such as choosing an orientation for the map, a position for the tree, and the placement of the labels. Several existing tools support drawing geophylogenies but we suspect that in practice many drawings are made “by hand”. The tools **GenGIS** by Parks et al. [23, 22], a tool by Page [20], and the R-package **phytools** by Revell [24] can generate side-by-side drawings with external labeling. The former two try to minimize leader crossings by testing random leaf orders and by rotating



■ **Figure 2** Side-by-side drawings of geophylogenies from the literature.

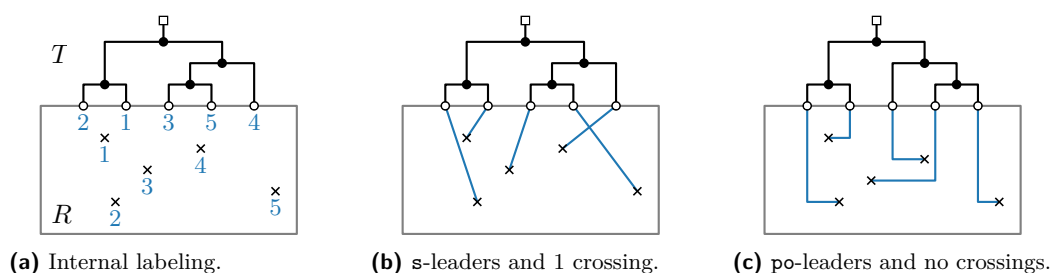
the phylogenetic tree around the map; Revell uses a greedy algorithm to minimize leader crossings. The R package `phylogeo` by Charlop-Powers and Brady [8] uses internal labeling via colors. Unfortunately, none of the articles describing these tools formally defines a quality measure being optimized or studies the underlying combinatorial optimization problem from an algorithmic perspective. In this paper, we introduce a simple combinatorial definition for side-by-side drawings of geophylogenies and propose several quality measures.

Labeling Geophylogenies

Following standard map-labeling terminology, *internal labeling* places the labels inside or in the direct vicinity of a feature; *external labeling* [5] places the labels in the margin next to the map and a label is connected to the corresponding feature with a *leader*. An *s-leader* is drawn using a single (straight) line segment as in Figures 2b and 3b. Alternatively, a *po-leader* (for: parallel, orthogonal) consists of a horizontal segment at the site and a vertical segment at the leaf; see Figure 3c. In the literature, we have only encountered s-leaders in geophylogeny drawings, but argue below that po-leaders should be considered. In a user study on external labeling, Barth, Gemsa, Niedermann, and Nöllenburg [1] showed that s-leaders perform well when users are asked to associate sites with their labels and vice versa, but that po-leaders (and “diagonal, orthogonal” do-leaders) are among the aesthetic preferences.

For internal labeling, a common optimization approach is to place the most labels possible such that none overlap; see Neyer [18] for a survey on this topic. Existing algorithms can be applied to label the sites in a geophylogeny drawing and it is geometrically straight-forward to place the labels for the leaves of T . However, a map reader must also be aided in associating the sites on the map with the leaves at the border based on these labels (and potentially colors). Consider the drawing in Figure 1c, which uses color-based internal labeling: the three kiwi species *A. australis*, *A. rowi*, and *A. mantelli* occur in this order from South to North. When using internal labeling, we would thus prefer, if possible, to have the three species in this order in the tree as well – as opposed to their order in Figure 1a.

External labeling styles conventionally forbid crossing the leaders as such crossings could be visually confusing (cf. Figure 2b). Often the total length of leaders is minimized given this constraint; see the survey by Bekos, Niedermann, and Nöllenburg [5]. If one allows a many-to-one correspondence between sites and labels, the literature typically seeks a drawing



■ **Figure 3** We place T above R and use either internal or external labeling to show the mapping between P and $L(T)$. Figures (b) and (c) minimize the number of crossings for their leader type. Note the difference in embedding of T and that not all permutations of leaves are possible.

that minimizes the number of crossings between the leaders, and this is NP-hard [17]. The problem remains NP-hard even when leaders can share segments, so-called hyper-leaders [2]. Even though our drawings of geophylogenies have a one-to-one correspondence, the planarity constraint on the tree restricts which leaf orders are possible and it is not always possible to have crossing-free leaders in a geophylogeny. In order to obtain a drawing with low visual complexity, our task is thus to find a leaf order that minimizes the number of leader crossings.

Results and Contribution

We formalize several graph visualization problems in the context of drawing geophylogenies. We propose quality measures for drawings with internal labeling and show that optimal solutions can be computed in quadratic time (Section 3). For external labeling (Section 4), we prove that although crossing minimization of s- and po-leaders is NP-hard in general, it is possible to check in polynomial time if a crossing-free drawing exists and to solve a certain class of instances efficiently in practice. Furthermore, we introduce an integer linear program (ILP) and several heuristics for crossing minimisation. We evaluate these solutions on synthetic and real-world examples and find that the ILP can solve realistic instances optimally in a matter of seconds and that the heuristics, which run in a fraction of a second, are often (near-)optimal as well (Section 5). We close the paper with a discussion and open problems; in particular, we point out further similarities between problems with geophylogeny drawings and with external labeling.

A longer version of this paper containing all proofs is available on arXiv [16]. Furthermore, implementations of the algorithms and the experiments are available online at github.com/joklawitter/geophylo.

2 Definitions and Notation

For a phylogenetic tree T , let $V(T)$ be its vertex set, $E(T)$ its edge set, $L(T)$ its leaves, and $I(T)$ its internal vertices. As size of an instance we let $n = |L(T)|$ be the number of leaves. Let $T(v)$ be the subtree rooted at v and $n(v) = |L(T(v))|$.

A map R is an axis-aligned rectangle and a site is a point on R . A geophylogeny G consists of a phylogenetic tree T , a map R , a set of points P on R , as well as a 1-to-1 mapping between $L(T) = \{\ell_1, \dots, \ell_n\}$ and $P = \{p_1, \dots, p_n\}$ so that without loss of generality the mapping is given by the indices.

We define a drawing Γ of G as consisting of drawings of R and T in the plane with the following properties (see Figure 3). We assume that T is always drawn at a fixed position above R such that the leaves of T lie at evenly spaced positions on the upper boundary

of R . Furthermore, we require that T is drawn *downward planar*, that is, all edges of T point downwards from the root towards the leaves, and no edges of T cross. (In our examples we draw T as a “rectangular cladogram”, but the exact drawing style is irrelevant given downward planarity.) The points of P are marked on R and the drawing uses either internal labeling as in Figure 3a or external labeling with s- or po-leaders as in Figures 3b and 3c. For drawings with external labeling, we use s_i to denote the leader that connects ℓ_i and p_i .

Since the tree is drawn without crossings and the sites have fixed locations, the only combinatorial freedom in the drawing Γ is the embedding of T , i.e. which child is to the left and which is to the right. Furthermore, since we fixed the relative positions of the map and the leaves, note that there is also no “non-combinatorial” freedom. Hence, an embedding of T corresponds one-to-one with a left-to-right order of $L(T)$ and we call this the *leaf order* π of Γ . For example, if a leaf ℓ_i is at position 4 in Γ , then $\pi(\ell_i) = 4$. Further, let $x(v)$ denote the x-coordinate of a site or leaf v of T in Γ .

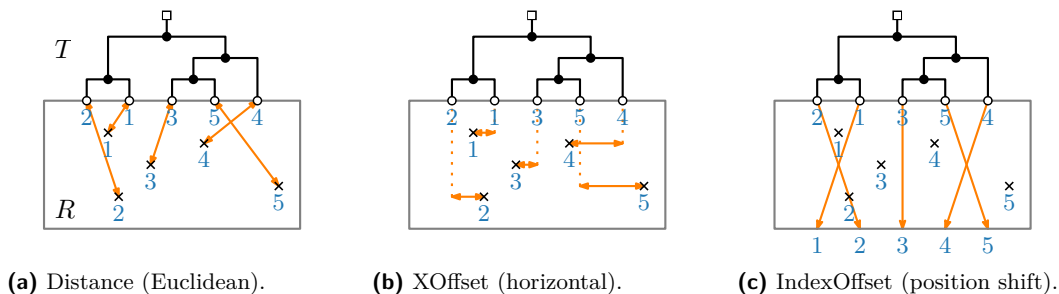
3 Geophylogenies with Internal Labeling

A good order of the leaves is crucial for internal labeling, since it can help the reader associate between $L(T)$ and P . It is in general not obvious how to determine which leaf order is best for this purpose; we propose three quality measures and a general class of measures that subsume them. Any measure in this class can be efficiently optimized by the algorithm described below. In practice one can easily try several quality measures and pick whichever suits the particular drawing; a user study of practical readability could also be fruitful.

3.1 Quality Measures

When visually searching for the site p_i corresponding to a leaf ℓ_i (or the opposite direction), it seems beneficial if ℓ_i and p_i are close together. Our first quality measure, *Distance*, sums the Euclidean distances over all pairs (p_i, ℓ_i) . Since the tree organizes the leaves from left to right, it might be better to consider only the horizontal distances, i.e. $\sum_{i=1}^n |x(p_i) - x(\ell_i)|$, which we call *XOffset*. Finally, instead of the geometric offset, *IndexOffset* considers how much the leaf order permutes the geographic left-to-right order of the sites. Assuming without loss of generality that the sites are indexed from left to right, we sum how many places each leaf ℓ_i is away from leaf position i , i.e. $\sum_{i=1}^n |\pi(\ell_i) - i|$. See Figure 4.

These measures have in common that they sum over some “quality” of the leaves, where the quality of a leaf depends only on its own position and that of the sites (but not the other leaves). We call such quality measures *leaf additive*. Unfortunately not all sensible quality measures are leaf additive (such as for example the number of inversions in π).



■ **Figure 4** Orange arrows indicate what the three quality measures for internal labeling consider.

3.2 Algorithm for Leaf-Additive Quality Measures

Let $f: L(T) \times \{1, \dots, n\} \rightarrow \mathbb{R}$ be a quality measure for placing one particular leaf at a particular position; the location of the sites is constant for a given instance, so we do not consider it an argument of f . This uniquely defines a leaf additive objective function on drawings by summing over the leaves; assume w.l.o.g. that we want to minimize this sum.

Now we naturally lift f to inner vertices of T by taking the sum over leaves in the subtree rooted at that vertex – in the best embedding of that subtree. More concretely, note that any drawing places the leaves of any subtree at consecutive positions and they take up a fixed width regardless of the embedding. Let $F(v, i)$ be the minimum, taken over all embeddings of $T(v)$ and assuming the leftmost leaf is placed at position i , of the sum of quality of the leaves of $T(v)$. Then by definition the optimal objective value for the entire instance is $F(w, 1)$, where w is the root of T .

► **Theorem 1.** *Let G be a geophylogeny on n taxa and let f be a leaf additive objective function. A drawing that minimizes (or maximizes) f can be computed in $\mathcal{O}(n^2)$ time.*

Proof. For a vertex v with children x and y , we observe the following equality, since the embedding has only two ways of ordering them and those subtrees are then independent.

$$F(v, i) = \min\{ F(x, i) + F(y, i + n(x)), F(y, i) + F(x, i + n(y)) \} \quad (1)$$

Using dynamic programming on F allows us to calculate $F(w, 1)$ in $\mathcal{O}(n^2)$ time and space, since there are $2n$ vertices, n possible leaf positions, and Equation (1) can be evaluated in constant time by precomputing all $n(v)$. The optimal embedding of T can be traced back through the dynamic programming table in the same runtime. ◀

Note that we can still define leaf additive quality measures when P contains regions (rather than just points) as in Figure 1. For example, instead of considering the distance between ℓ_i and p_i , we could consider the smallest distance between ℓ_i and any point in the region p_i .

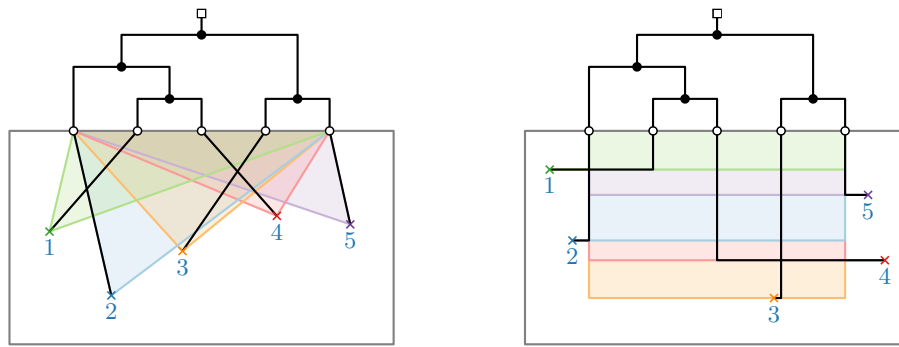
With the above algorithm, we can restrict leaves and subtrees to be in a certain position or a range of positions, simply by marking all other positions as prohibitively expensive in F ; the rotation of an inner vertex can also be fixed by considering only the corresponding term of Equation (1). This can be used if there is a conventional order for some taxa or to ensure that an outgroup-taxon is placed at the leftmost or rightmost position. Furthermore, this enables an interactive editing experience where a designer can inspect the initial optimized drawing and receive re-optimized versions based on their feedback – for example “put the leaves for the sea lions only where there is water on the edge of the map”. (This is leaf additive.)

4 Geophylogenies with External Labeling

For external labeling, the optimization goal is to embed the tree such that the number of crossings between leaders is minimized. Unless otherwise stated, we use **s**-leaders.

For brevity, we omit proofs¹ of following foundational complexity results and move on to more practical algorithms.

¹ Proofs are available in the long version of this paper [16].



(a) A geometry-free instance for **s**-leaders: no site lies inside the **s**-area of another site.

(b) A geometry-free instance for **po**-leaders: no site lies inside the **po**-area of another site.

■ **Figure 5** In a geometry-free instance the leaf order π fully determines if any two leaders cross.

► **Proposition 2.** *Given a geophylogeny G and an integer k , it is NP-hard to decide, for both **s**- and **po**-leaders, if G admits a drawing with external labels and at most k leader crossings.*

► **Proposition 3.** *Given a geophylogeny G on n taxa, it can be decided in $\mathcal{O}(n^6)$ time, for both **s**- and **po**-leaders, whether G admits a drawing with external labels and no leader crossings.*

4.1 Geometric Structure and Geometry-Free Instances

We start by making some observations about the structure of geophylogeny drawings. This leads to an $\mathcal{O}(n \log n)$ -time algorithm for crossing minimization on a particular class of “geometry-free” instances and forms the basis for our ILP.

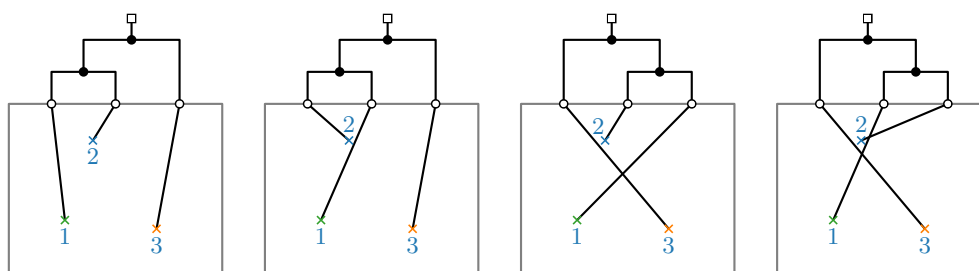
Let B be the line segment between leaf position 1 (left) and leaf position n (right); let the **s**-area of a site p_i be the triangle spanned by p_i and B . Note that the leader s_i lies within this triangle in any drawing. Now consider two sites p_i and p_j that lie outside each other’s **s**-area. Independently of the embedding of the tree, s_i always passes p_j on the same side: see Figure 5 where, for example, s_2 passes left of p_4 in any drawing. As a result, if p_i lies left of p_j , then s_i and s_j cross if and only if the leaf ℓ_i is positioned right of the leaf ℓ_j (cf. Figure 5). The case where p_i is right of p_j is flipped. We call such a pair (p_i, p_j) *geometry free* since purely the *order* of the corresponding leaves suffices to recognize if their leaders cross: the precise geometry of the leaf positions is irrelevant.

Conversely, consider a site p_k that lies inside the **s**-area of p_i . Whether the leaders s_i and s_k cross depends on the placement of the leaves ℓ_i and ℓ_k in a more complicated way than just their relative order: s_i might pass left or right of p_k . In this case, we call p_i *undecided* with respect to p_k . See Figure 6, where p_1 is undecided with respect to p_2 .

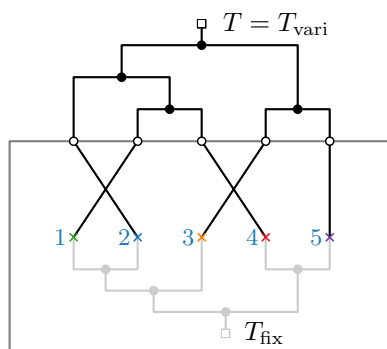
We call a geophylogeny *geometry free* if all pairs of sites are geometry free. Such instances are not entirely implausible: for example, researchers may have taken their samples along a coastline, a river, or a valley, in which case the sites may lie relatively close to a line. Orienting the map such that this line is horizontal could result in a geometry-free instance. Furthermore, unless two sites share an x-coordinate, increasing the vertical distance between the map and the tree eventually results in a geometry-free drawing for **s**-leaders; however, the required distance might be impractically large.

Concerning **po**-leaders, we can analogously define the **po**-area of a site (see Figure 5b).

► **Theorem 4.** *Given a geometry-free geophylogeny G on n taxa, a drawing with the minimum number of leader crossings can be found in $\mathcal{O}(n \log n)$ time, for both **s**- and **po**-leaders.*



■ **Figure 6** Drawings of the same geophylogeny with different leaf orders. Whether s_1 and s_2 cross depends on the position of ℓ_1 and ℓ_2 , whereas s_1 and s_3 cross if and only if ℓ_3 is left of ℓ_1 . We call the pair (p_1, p_2) *undecided* and the pair (p_1, p_3) *geometry-free*.



■ **Figure 7** A geometry-free geophylogeny and a one-sided tanglegram $(T_{\text{fix}}, T_{\text{vari}})$ that have the same combinatorics (in terms of leader crossings) as the two geometry-free instances in Figure 5.

Proof. We transform G into a so-called *one-sided tanglegram* $(T_{\text{fix}}, T_{\text{vari}})$ that is equivalent in terms of crossings; see Figure 7. In a *tanglegram* [11] two phylogenetic trees on the same taxa are drawn planar opposite each other and the matching taxa are connected with straight line segments; the goal is to find leaf orders that minimize the number of crossings. In a one-sided tanglegram, the leaf order for one tree is given and fixed.

We take the sites P as the leaves of T_{fix} and embed the tree so that the points are ordered from left to right; the topology of T_{fix} is arbitrary. As the tree T_{vari} with variable embedding, we take the phylogenetic tree T . Since G is geometry-free, the crossings in the tanglegram correspond one-to-one with those in the geophylogeny drawing with the same embedding.

The number of crossings of $(T_{\text{fix}}, T_{\text{vari}})$ can be minimized in $\mathcal{O}(n \log n)$ time using an algorithm of Fernau et al. [11]: the resulting leaf order for T_{vari} then also minimizes the number of leader crossings in Γ . ◀

4.2 Optimal Drawings with Integer Linear Programming

For the following ILP, we consider an arbitrary embedding of the tree as *neutral* and describe all embeddings in terms of which internal vertices of T are rotated with respect to this neutral embedding, i.e. for which internal vertices to swap the left-to-right order of their two children. For two sites p_i and p_j , we use $p_i \prec p_j$ to denote that ℓ_i is left of ℓ_j in the neutral embedding. Let U be the set of undecided pairs, that is, all ordered pairs (p, q) where q lies inside the \mathbf{s} -area of p ; note that these are ordered pairs.

Variables and Objective Function

$\rho_i \in \{0, 1\} \forall i \in I(T)$. Do we rotate internal vertex i (1) or keep its neutral embedding (0)?

Note that rotating the lowest common ancestor of ℓ_i and ℓ_j is the only way to flip their order, so for convenience we write ρ_{ij} to mean $\rho_{\text{lca}(i,j)}$.

$d_{pq} \in \{0, 1\} \forall (p, q) \in U$. For each undecided pair (p, q) : should p 's leader pass to the left (0) or to the right (1) of site q ? (This is well-defined since the pair is undecided.)

$\chi_{pq} \in \{0, 1\} \forall p, q \in P, p < q$. For each set of two sites: are the leaders of p and q *allowed* to cross? There is no requirement that noncrossing pairs have $\chi_{pq} = 0$, but that will be the case in an optimal solution.

To minimise the number of crossings, minimize the sum over all χ_{pq} .

Constraints

We handle geometry-free pairs and undecided pairs separately.

Consider a geometry-free pair of sites: if the leaders cross in the neutral embedding, we must either allow this, or rotate the lowest common ancestor. Conversely, if they do not cross neutrally, yet we rotate the lowest common ancestor, then we must allow their leaders to cross. Call these sets of pairs F_{rotate} and F_{keep} respectively, for how to prevent the crossing.

$$\chi_{ij} + \rho_{ij} \geq 1 \quad \forall (i, j) \in F_{\text{rotate}}; \quad \chi_{ij} - \rho_{ij} \geq 0 \quad \forall (i, j) \in F_{\text{keep}} \quad (2)$$

For undecided pairs (p, q) , a three-way case distinction on $[p \prec q]$, ρ_{pq} and d_{pq} reveals the following geometry: pairs with $p \prec q$ have crossing leaders if and only if $\rho_{pq} + d_{pq} = 1$; pairs with $p \succ q$ have crossing leaders if and only if $\rho_{pq} + d_{pq} \neq 1$. Recall that we do not force χ to be zero if there is no intersection, only that it is 1 if there *is* an intersection; we implement these conditions in the ILP as follows. Let $U_{\text{left}} \subseteq U$ be the undecided pairs with $p \prec q$.

$$\rho_{pq} - d_{pq} \leq \chi_{pq} \quad \forall (p, q) \in U_{\text{left}}; \quad d_{pq} - \rho_{pq} \leq \chi_{pq} \quad \forall (p, q) \in U_{\text{left}} \quad (3)$$

Conversely, let $U_{\text{right}} \subseteq U$ be the undecided pairs with $p \succ q$.

$$\rho_{pq} + d_{pq} - 1 \leq \chi_{pq} \quad \forall (p, q) \in U_{\text{right}}; \quad 1 - \rho_{pq} - d_{pq} \leq \chi_{pq} \quad \forall (p, q) \in U_{\text{right}} \quad (4)$$

Finally, we must ensure that each leader s_i respects the d variables: the \mathbf{s} -leader from p_i to ℓ_i must pass by each other site in the \mathbf{s} -area on the correct side. This does not affect geometry-free pairs, but we must constrain the leaf placement for undecided pairs.

Observe that the ρ variables together fix the leaf order, since they fix the embedding of T . Let $L_i(\rho)$ be the function that gives the x-coordinate of ℓ_i given the ρ variables. Note that L_i is linear in each of the ρ variables: rotating an ancestor of ℓ_i shifts its leaf location by a particular constant, and rotating a non-ancestor does not affect it.

For an undecided pair (p_i, p_j) , let $x^*(i, j)$ be the x-coordinate of where the ray from p_i through p_j intersects the top of the map and note that this is a constant. If $d_{ij} = 0$, then ℓ_i must be to the left of this intersection; if $d_{ij} = 1$, it must be to the right. We model this in the ILP with two constraints and the *big-M method*, where we can set $M = n$.

$$L_i(\rho) - d_{ij}M \leq x^*(i, j), \quad L_i(\rho) + (1 - d_{ij})M \geq x^*(i, j); \quad \forall (p_i, p_j) \in U \quad (5)$$

The number of variables and constraints in the ILP are both quadratic in n .

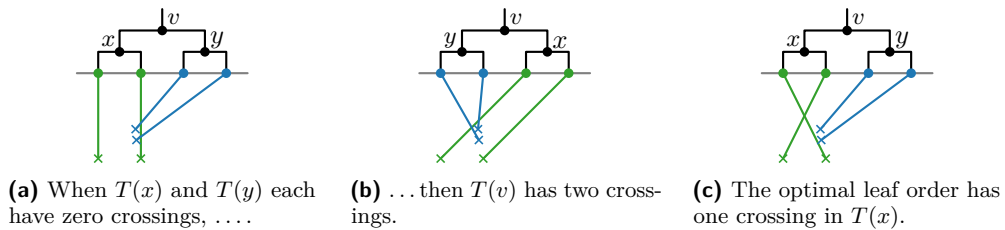
4.3 Heuristics

Since the ILP from the previous section can be slow in the worst case and requires advanced solver software, we now suggest a number of heuristics.

Bottom-Up. First, we use a dynamic program similar to the one in Section 3 and commit to an embedding for each subtree while going up the tree. At this point we note that counting the number of crossings is not a leaf additive objective function in the sense of Section 3. However, Equation (1) does enable us to introduce an additional cost based on where an entire subtree is placed and where its sibling subtree is placed – just not minimized over the embedding of these subtrees. More precisely, for an inner vertex v of T with children x and y , let $C(x, y, i)$ be the number of crossings between $T(x)$ and $T(y)$ when placed starting at position i and $i + n(x)$ respectively; this can be computed in $\mathcal{O}(n(v)^2)$ time. Note that this ignores any crossings with leaders from other subtrees. With base case $H(\ell, i) = 0$ for every leaf ℓ , we use

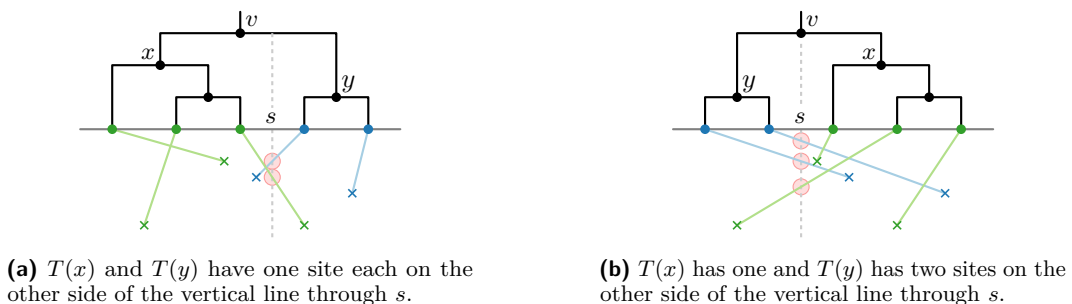
$$H(v, i) = \min\{ H(x, i) + H(y, i + n(x)) + C(x, y, i), H(y, i) + H(x, i + n(y)) + C(y, x, i) \}$$

to pick a rotation of $T(v)$. Since this can be evaluated in $\mathcal{O}(n^2)$ time, the heuristic runs in $\mathcal{O}(n^4)$ time. In the example in Figure 8 this does not minimize the total number of crossings.



■ **Figure 8** The bottom-up heuristic is not always optimal.

Top-Down. The second heuristic traverses T from top to bottom (i.e. in pre-order) and chooses a rotation for each inner vertex v based on how many leaders would cross the vertical line between the two subtrees of v ; see Figure 9. More precisely, suppose that $T(v)$ has its leftmost leaf at position i based on the rotations of the vertices above v . For x and y the children of v , consider the rotation of v where $T(x)$ is placed starting at position i and $T(y)$ is placed starting at position $i + n(x)$. Let s be the x-coordinate in the middle between the last leaf of $T(x)$ and the first leaf $T(y)$. We compute the number of leaders of $T(v)$ that cross the vertical line at s and for the reverse rotation of v ; the smaller result is chosen and the rotation fixed. This procedure considers each site at most $\mathcal{O}(n)$ times and thus runs in $\mathcal{O}(n^2)$ time.



■ **Figure 9** The top-down heuristic tries both rotations of v and here would pick (a).

Leaf-Additive Dynamic Programming. Thirdly, we could optimize any of the quality measures for interior labeling (Section 3). These measures produce generally sensible leaf orders in quadratic time and we may expect the number of leader crossings to be low.

Greedy (Hill Climbing). Finally, we consider a hill climbing algorithm that, starting from some leaf order, greedily performs rotations that improve the number of crossings. This could start from a random leaf order, a hand-made one, or from any of the other heuristics. Evaluating a rotation can be done in $\mathcal{O}(n^2)$ time and thus one round through all vertices runs in $\mathcal{O}(n^3)$ time.

5 Experimental Evaluation

This section is based on our implementation of the ILP and the heuristics. The code is available online at github.com/joklawitter/geophylo, and data from the corresponding authors upon request.

5.1 Test Data

We use three procedures to generate random instances. For each type and with 10 to 100 taxa (in increments of 5), we generated 10 instances; we call these the *synthetic instances*. We stop at 100 since geophylogeny drawings with more taxa are rarely well-readable.

Uniform. Place n sites on the map uniformly at random. Generate the phylogenetic tree by repeating this merging procedure. Pick an unmerged site or a merged subtree uniformly at random, then pick a second with probability distributed by inverse distance to the first, and merge them; as position of a subtree, we take the median coordinate on both axis.

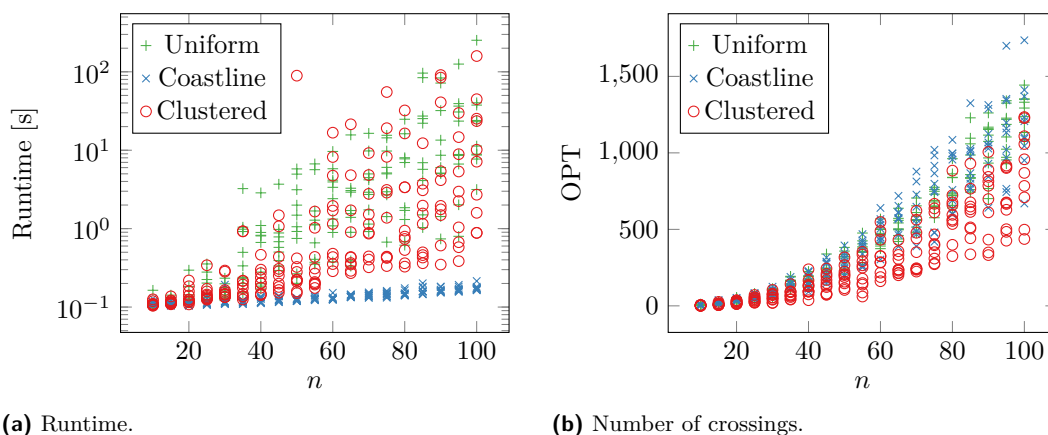
Coastline. Initially place all sites equidistantly on a horizontal line, then slightly perturb the x-coordinates. Next, starting at the central site and going outwards, change the y-coordinate of each site randomly (up to 1.5 times the horizontal distance) from the y-coordinate of the previous site. Construct the tree as before.

Clustered. These instances group multiple taxa into clusters. First a uniformly random number of sites between three and ten is allocated for a cluster and its center is placed at a uniformly random point on the map. Then for each cluster, we place sites randomly in a disk around the center with size proportional to the cluster size. Construct T as before, but first for each cluster separately and only then for the whole instance.

In addition, we consider three real world instances derived from published drawings. **Fish** is a 14-taxon geophylogeny by Williams and Johnson [28]. **Lizards** is 20-taxon geophylogeny by Jauss et al. [13], where the sites are mostly horizontally dispersed (see Figure 2b). **Frogs** is a 64-taxon geophylogeny by Ellepola et al. [10], where the sites are rather randomly dispersed on the map; the published drawing with s-leaders has over 680 leader crossings.

5.2 Experimental Results

The ILP is fairly quick. Our implementation uses a Python script to generate the ILP instance and Gurobi 10 to solve it; we ran the experiments on a 10-core Apple M1 Max processor. As expected, we observe that the runtime is exponential in n , but only moderately so (Figure 10). Instances with up to about 50 taxa can usually be solved optimally within a second, but for Clustered and Uniform instances the ILP starts to get slow at about 100 taxa. We note that geophylogenies with over 100 taxa should probably not be drawn with external labeling: for example, the Frogs instance can be drawn optimally by the ILP in



(a) Runtime.

(b) Number of crossings.

■ **Figure 10** Computing optimal drawings with the ILP.

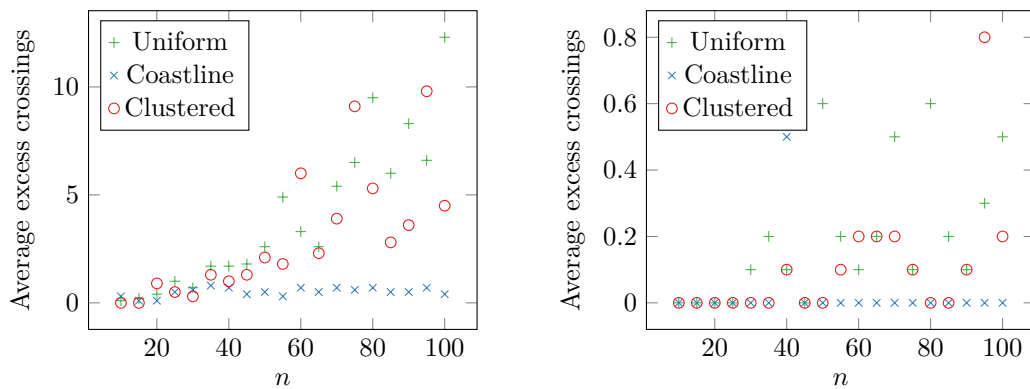
about 0.5 s, but even though this improves the number of crossings from the published 680 to the optimal 609, the drawing is so messy as to be unreadable (Figure 12b). We further observe that Coastline instances are solved trivially fast, since with fewer undecided pairs the ILP is smaller and presumably easier to solve.

The synthetic instances have a superlinear number of crossings. The Clustered instances can be drawn with significantly fewer crossings than Uniform: this matches our expectation, as by construction there is more correlation between the phylogenetic tree and the geography of the sites. More surprisingly we find that the Coastline instances require many crossings. We may have made them too noisy, but this does warn of the generally quadratic growth in number of crossings, which makes external labeling unsuitable for large geophylogenies unless the geographic correlation is exceptionally good.

The heuristics run instantly and Greedy is often optimal. The heuristics are implemented in single-threaded Java code. Bottom-Up, Top-Down and Leaf-Additive all run instantly, and even the Greedy hill climber runs in a fraction of a second. Of the first three heuristics, Bottom-Up consistently achieves the best results for both \mathbf{s} - and \mathbf{po} -leaders. Comparing the best solution by these heuristics with the optimal drawing (Figure 11), we observe that the number crossings in excess of the optimum increases with the number of taxa, in particular for Uniform and Clustered instances; Coastline instances are always drawn close to optimally by at least one heuristic. The Greedy hill climber often improves this to an optimal solution.

For the number of crossings, \mathbf{po} -leaders are promising. In addition to \mathbf{s} -leaders, our implementation of the heuristics can handle \mathbf{po} -leaders. (The ILP cannot.) Our heuristics require on average only about 73% as many crossings when using \mathbf{po} -leaders compared to \mathbf{s} -leaders (55% for Coastline instances); the Lizard example in Figure 2b requires 11 \mathbf{s} -leader crossings but only 2 \mathbf{po} -leader crossings. We therefore propose that \mathbf{po} -leaders deserve more attention from the phylogenetic community.

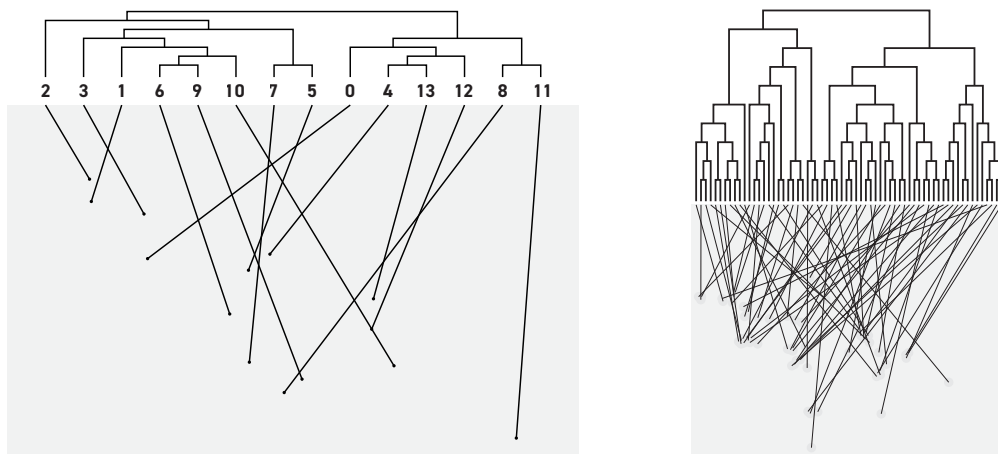
Algorithmic recommendations. Our results show that the ILP is a good choice for geophylogeny drawings with external labeling. If no solver is at hand or it is technically challenging to set up (for example when making an app that runs locally in a user's web browser), then the heuristics offer an effective and efficient alternative, especially Bottom-Up and Greedy.



(a) Best heuristic without Greedy.

(b) Best after Greedy postprocessing.

Figure 11 Number of crossings made by the best heuristic minus the number of crossings in the optimal drawing, averaged over 10 random instances per value of n .



(a) Drawing of **Fish** with 17 crossings.

(b) Drawing of **Frogs** with 609 crossings.

Figure 12 Crossing-optimal drawings of Fish and Frogs with s -leaders.

For the Fish instance, for example, we found that the drawing with s -leaders and 17 crossings in Figure 12a is a good alternative to the internal labeling used in the published drawing [28]. However, for instances without a clear structure or with many crossings, it might be better to use internal labeling. Alternatively, the tree could be split like Tobler et al. [26], such that different subtrees are each shown with the map in separate drawings.

6 Discussion and Open Problems

In this paper, we have shown that drawings of geophylogenies can be approached theoretically and practically as a problem of algorithmic map labeling. We formally defined a drawing style for geophylogenies that uses either internal labeling with text or colors, or that uses external labeling with s / po -leaders. This allowed us to define optimization problems that can be tackled algorithmically. For drawings with internal labeling, we introduced a class of quality measures that can be optimized efficiently and even interactively constrained. In practice, designers can thus try different quality measures, pick their favorite, and make further

adjustments easily even for large instances. For external labeling, minimizing the number of leader crossings is NP-hard in general, but we provide multiple algorithmic approaches to solve this problem and demonstrated experimentally that they perform well in practice.

Even though we have provided a solid base of results, we feel the algorithmic study of geophylogeny drawings holds further promise by varying, for example, the type of leader used, the objective function, the composition of the drawing, or the nature of the phylogeny and the map. We finish this paper with several suggestions for future work.

One might consider *do*- and *pd*-leaders, which use a diagonal segment and can be aesthetically pleasing. We expect that some of our results (such as the NP-hardness of crossing minimization and the effectiveness of the heuristics) should hold for these leaders. The boundary labeling literature [5] studies even further types, such as *opo* and Bézier, and these might be more challenging to adapt.

For external labeling we have only considered the total number of crossings. If different colors are used for the leaders of different clades or if the drawing can be explored with an interactive tool, one might want to minimize the number of crossings within each clade (or a particular clade). Furthermore, one might optimize crossing angles. While we provided heuristics to minimize leader crossings, the development of approximation algorithms, which exist for other labeling problems [17, 3], could also be of interest.

Our model of a geophylogeny drawing can be expanded. One might allow the orientation of the map to be freely rotated, the extent of the map to be changed, or the leaves to be placed non-equidistantly. Optimizing over these additional freedoms poses new algorithmic challenges. Straying further from our model, some drawings in the literature have a circular tree around the map [21, 14]. (This is similar to contour labeling in the context of map labeling [19].) Also recall that Figure 1 has area features. Our quality measures for internal labeling are easily adapted to handle this, but (as is the case with general boundary labeling [4]) area features provide additional algorithmic challenges for external labeling. The literature contains many drawings where multiple taxa correspond to the same feature on the map [7], where we might want to look to many-to-one boundary labeling [17, 2]. Furthermore, one can consider non-binary phylogenetic trees and phylogenetic networks.

Lastly, we note that side-by-side drawings can also be used for a phylogenetic tree together with a diagram other than a map: Chen et al. [9] combine it with a scatter plot; Gehring et al. [12] even combine three things (phylogenetic tree, haplotype network, and map).

References

- 1 Lukas Barth, Andreas Gemsa, Benjamin Niedermann, and Martin Nöllenburg. On the readability of leaders in boundary labeling. *Information Visualization*, 18(1), 2019. doi:10.1177/1473871618799500.
- 2 Michael A. Bekos, Sabine Cornelsen, Martin Fink, Seok-Hee Hong, Michael Kaufmann, Martin Nöllenburg, Ignaz Rutter, and Antonios Symvonis. Many-to-one boundary labeling with backbones. *Journal of Graph Algorithms and Applications*, 19(3):779–816, 2015. doi:10.7155/jgaa.00379.
- 3 Michael A. Bekos, Michael Kaufmann, Dimitrios Papadopoulos, and Antonios Symvonis. Combining traditional map labeling with boundary labeling. In Ivana Cerná, Tibor Gyimóthy, Juraj Hromkovic, Keith G. Jeffery, Rastislav Královic, Marko Vukolic, and Stefan Wolf, editors, *SOFSEM 2011*, volume 6543 of *LNCS*, pages 111–122. Springer, 2011. doi:10.1007/978-3-642-18381-2_9.
- 4 Michael A. Bekos, Michael Kaufmann, Katerina Potika, and Antonios Symvonis. Area-feature boundary labeling. *The Computer Journal*, 53(6):827–841, 2010. doi:10.1093/comjnl/bxp087.

- 5 Michael A. Bekos, Benjamin Niedermann, and Martin Nöllenburg. External labeling techniques: A taxonomy and survey. *Computer Graphics Forum*, 38(3):833–860, 2019. doi:10.1111/cgf.13729.
- 6 R Alexander Bentley, William R Moritz, Damian J Ruck, and Michael J O'Brien. Evolution of initiation rites during the austronesian dispersal. *Science Progress*, 104(3):00368504211031364, 2021. doi:10.1177/00368504211031364.
- 7 Tyler K Chafin, Marlis R Douglas, Whitney JB Anthonysamy, Brian K Sullivan, James M Walker, James E Cordes, and Michael E Douglas. Taxonomic hypotheses and the biogeography of speciation in the tiger whiptail complex. *Frontiers*, 13(2), 2021. doi:10.21425/F5FBG49120.
- 8 Zachary Charlop-Powers and Sean F. Brady. phylogeo: an R package for geographic analysis and visualization of microbiome data. *Bioinformatics*, 31(17):2909–2911, 2015. doi:10.1093/bioinformatics/btv269.
- 9 Yi Chen, Lei Zhao, Huajing Teng, Chengmin Shi, Quansheng Liu, Jianxu Zhang, and Yaohua Zhang. Population genomics reveal rapid genetic differentiation in a recently invasive population of *rattus norvegicus*. *Frontiers in Zoology*, 18(1):6, 2021. doi:10.1186/s12983-021-00387-z.
- 10 Gajaba Ellepola, Jayampathi Herath, Kelum Manamendra-Arachchi, Nayana Wijayathilaka, Gayani Senevirathne, Rohan Pethiyagoda, and Madhava Meegaskumbura. Molecular species delimitation of shrub frogs of the genus *pseudophilautus* (anura, rhacophoridae). *PLOS ONE*, 16(10):1–17, 2021. doi:10.1371/journal.pone.0258594.
- 11 Henning Fernau, Michael Kaufmann, and Mathias Poths. Comparing trees via crossing minimization. *Journal of Computer and System Sciences*, 76(7):593–608, 2010. doi:10.1016/j.jcss.2009.10.014.
- 12 Philip-Sebastian Gehring, Maciej Pabijan, Jasmin E. Randrianirina, Frank Glaw, and Miguel Vences. The influence of riverine barriers on phylogeographic patterns of malagasy reed frogs (*heterixalus*). *Molecular Phylogenetics and Evolution*, 64(3):618–632, 2012. doi:10.1016/j.ympev.2012.05.018.
- 13 Robin-Tobias Jauss, Nadiné Solf, Sree Rohit Raj Kolora, Stefan Schaffer, Ronny Wolf, Klaus Henle, Uwe Fritz, and Martin Schlegel. Mitogenome evolution in the *lacerta viridis* complex (lacertidae, squamata) reveals phylogeny of diverging clades. *Systematics and Biodiversity*, 19(7):682–692, 2021. doi:10.1080/14772000.2021.1912205.
- 14 Monika Karmin, Rodrigo Flores, Lauri Saag, Georgi Hudjashov, Nicolas Brucato, Chelzie Crenna-Darusallam, Maximilian Larena, Phillip L Endicott, Mattias Jakobsson, J Stephen Lansing, Herawati Sudoyo, Matthew Leavesley, Mait Metspalu, François-Xavier Ricaut, and Murray P Cox. Episodes of Diversification and Isolation in Island Southeast Asian and Near Oceanian Male Lineages. *Molecular Biology and Evolution*, 39(3), 2022. doi:10.1093/molbev/msac045.
- 15 David M. Kidd and Xianhua Liu. geophylobuilder 1.0: an arcgis extension for creating “geophylogenies”. *Molecular Ecology Resources*, 8(1):88–91, 2008. doi:10.1111/j.1471-8286.2007.01925.x.
- 16 Jonathan Klawitter, Felix Klesen, Joris Y. Scholl, Thomas C. van Dijk, and Alexander Zaft. Visualizing geophylogenies – internal and external labeling with phylogenetic tree constraints. *CoRR*, abs/2306.17348, 2023. arXiv:2306.17348.
- 17 Chun-Cheng Lin, Hao-Jen Kao, and Hsu-Chun Yen. Many-to-one boundary labeling. *Journal of Graph Algorithms and Applications*, 12(3):319–356, 2008. doi:10.7155/jgaa.00169.
- 18 Gabriele Neyer. Map labeling with application to graph drawing. In Michael Kaufmann and Dorothea Wagner, editors, *Drawing Graphs: Methods and Models*, pages 247–273. Springer, 2001. doi:10.1007/3-540-44969-8_10.
- 19 Benjamin Niedermann, Martin Nöllenburg, and Ignaz Rutter. Radial contour labeling with straight leaders. In Daniel Weiskopf, Yingcai Wu, and Tim Dwyer, editors, *IEEE Pacific Visualization Symposium*, pages 295–304. IEEE Computer Society, 2017. doi:10.1109/PACIFICVIS.2017.8031608.

- 20 Roderic Page. Visualising geophylogenies in web maps using geojson. *PLOS Currents*, 7, 2015. doi:10.1371/currents.tol.8f3c6526c49b136b98ec28e00b570a1e.
- 21 Da Pan, Boyang Shi, Shiyu Du, Tianyu Gu, Ruxiao Wang, Yuhui Xing, Zhan Zhang, Jiajia Chen, Neil Cumberlidge, and Hongying Sun. Mitogenome phylogeny reveals Indochina Peninsula origin and spatiotemporal diversification of freshwater crabs (Potamidae: Potamiscinae) in China. *Cladistics*, 38(1):1–12, 2022. doi:10.1111/c1a.12475.
- 22 Donovan H. Parks, Timothy Mankowski, Somayeh Zangoeei, Michael S. Porter, David G. Armanini, Donald J. Baird, Morgan G. I. Langille, and Robert G. Beiko. GenGIS 2: geospatial analysis of traditional and genetic biodiversity, with new gradient algorithms and an extensible plugin framework. *PLoS ONE*, 8(7):1–10, 2013. doi:10.1371/journal.pone.0069885.
- 23 Donovan H. Parks, Michael Porter, Sylvia Churcher, Suwen Wang, Christian Blouin, Jacqueline Whalley, Stephen Brooks, and Robert G. Beiko. GenGIS: A geospatial information system for genomic data. *Genome Research*, 19(10):1896–1904, 2009. doi:10.1101/gr.095612.109.
- 24 Liam J. Revell. phytools: an r package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution*, 3(2):217–223, 2012. doi:10.1111/j.2041-210X.2011.00169.x.
- 25 Mike Steel. *Phylogeny: Discrete and Random Processes in Evolution*. Society for Industrial and Applied Mathematics, 2016. doi:10.1137/1.9781611974485.
- 26 Ray Tobler, Adam Rohrlach, Julien Soubrier, Pere Bover, Bastien Llamas, Jonathan Tuke, Nigel Bean, Ali Abdullah-Highfold, Shane Agius, Amy O’Donoghue, Isabel O’Loughlin, Peter Sutton, Fran Zilio, Keryn Walshe, Alan N. Williams, Chris S M Turney, Matthew Williams, Stephen M Richards, Robert J Mitchell, Emma Kowal, John R Stephen, Lesley Williams, Wolfgang Haak, and Alan Cooper. Aboriginal mitogenomes reveal 50,000 years of regionalism in australia. *Nature*, 544(7649):180–184, 2017. doi:10.1038/nature21416.
- 27 Jason T. Weir, Oliver Haddrath, Hugh A. Robertson, Rogan M. Colbourne, and Allan J. Baker. Explosive ice age diversification of kiwi. *Proceedings of the National Academy of Sciences*, 113(38):E5580–E5587, 2016. doi:10.1073/pnas.1603795113.
- 28 Trevor J. Williams and Jerald B. Johnson. History predicts contemporary community diversity within a biogeographic province of freshwater fish. *Journal of Biogeography*, 49(5):809–821, 2022. doi:10.1111/jbi.14316.
- 29 Xuhua Xia. Pgt: Visualizing temporal and spatial biogeographic patterns. *Global Ecology and Biogeography*, 28(8):1195–1199, 2019. doi:10.1111/geb.12914.

Map Reproducibility in Geoscientific Publications: An Exploratory Study

Eftychia Koukouraki¹ ✉ 

Institute for Geoinformatics, University of Münster, Germany

Christian Kray ✉ 

Institute for Geoinformatics, University of Münster, Germany

Abstract

Reproducibility is a core element of the scientific method. In the Geosciences, the insights derived from geodata are frequently communicated through maps, and the computational methods to create these maps vary in their ease of reproduction. In this paper, we present the results from a study where we tried to reproduce the maps included in geoscientific publications. Following a systematic approach, we collected 27 candidate papers and in four cases, we were able to successfully reproduce the maps they contained. We report on the approach we applied, the issues we encountered and the insights we gained while attempting to reproduce the maps. In addition, we provide an initial set of criteria to assess the success of a map reproduction attempt. We also propose some guidelines for improving map reproducibility in geoscientific publications. Our work sheds a light on the current state of map reproducibility in geoscientific papers and can benefit researchers interested in publishing maps in a more reproducible way.

2012 ACM Subject Classification Human-centered computing → Geographic visualization; Applied computing → Cartography

Keywords and phrases Reproducible Research, Reproduction Assessment, Map Making, Cartography

Digital Object Identifier 10.4230/LIPIcs.GIScience.2023.6

Supplementary Material *Collection (Papers):* <https://doi.org/10.17605/OSF.IO/P5V2Z>

Acknowledgements We would like to thank the authors who took the time to respond to our queries and helped us to overcome issues we faced in reproducing maps and the anonymous reviewers for their constructive feedback.

1 Introduction

The reproducibility of research results is a critical aspect across all scientific disciplines, and the domain of Geosciences is no exception. It is widely accepted by the scientific community that the ability to reproduce results by other working groups enhances the trust and the reliability of the respective research. In an effort to clarify the different existing terminologies for *reproducibility* and *reproducible research*, Barba [1] defines reproducible research as the case when “authors provide all the necessary data and the computer codes to run the analysis again, re-creating the results”, while scientific *replication* can be achieved when a study produces the same results using different methods or different data. This distinction is also known as the Claerbout [3]/ Donoho [4]/ Peng [28] convention.

In recent years, reproducibility has rapidly gained relevance in different areas of Geosciences. In the *Forum on Reproducibility and Replicability in Geography* introduced by Goodchild et al. [8], the discussed topics range from the theoretical dimension of reproducibility and replicability in Geography [34], to more tangible matters, such as the review of current technological solutions [25] in this context.

¹ Corresponding Author



6:2 Map Reproducibility in Geoscientific Publications: An Exploratory Study

Figures are key elements of geoscientific publications and play an important role in the dissemination of scientific findings. Konkol and Kray [15] focused on the role of figures as a means of communicating research results and conducted a survey to identify the nature of the incorporated figures (e.g., maps, time series, histograms, etc.), as well as the frequency that they are used in geoscientific papers. The survey showed that maps were most frequently mentioned, which seems reasonable for this domain. Although there are several definitions for what a map is, most of them agree that maps are abstract representations of the real world [18], in which certain attributes of reality are highlighted, while others are deemphasised or completely left out. In line with fundamental scientific principles and the core role maps play in communicating societally relevant scientific outcomes (e.g., climate change, epidemiological spread), it is essential to be able to reproduce them. This capacity to reproduce results builds trust [10], which is essential for the uptake of scientific outcomes in society.

The aim of the research reported in this paper is thus to investigate the reproducibility of maps that are published as part of scientific publications. For this purpose, we screened 27 open access papers from the Geosciences and tried to reproduce the maps contained therein. In doing so, our objectives were (i) to investigate the current practice of creating maps for scientific publications, (ii) to assess the material availability, (iii) to analyse the complexities of map reproduction, and (iv) determine the degree of reproducibility of recently published maps in geoscientific articles. Our findings show that reproducing maps from recent papers is rarely easy, frequently requires extensive efforts and can even fail despite data and code being available. The fact that currently there are no well-defined success criteria for map reproduction further hampers it. Our main contributions are an initial set of criteria for assessing the success of map reproduction, insights into challenges arising during map reproduction and a set of guidelines for making map reproduction easier. These contributions pave the way for further research into map reproduction and can help researchers in making the maps they publish more reproducible.

The rest of the paper is structured as follows. Section 2 contextualises our work, reviews related studies and initiatives, and defines the term *reproducible map making*. In section 3, we motivate and describe the methodological workflow that we followed for the reproduction study. Section 4 presents the results of the study, including the obstacles that were encountered and the reproduced maps next to the original ones. Section 5 discusses the insights that were revealed and proposes a set of initial guidelines for making maps more reproducible. Section 6 summarises our key findings and outlines future work.

2 Background

In this section, we briefly summarise related work on map creation, open science practices, reproducibility in the Geosciences in general and in map production in particular.

2.1 Map Production

The International Cartographic Association (ICA) defines Cartography as the science, art, and technology of map making and map use [17]. Taylor [35] describes cartography in the context of Geographic Information Systems (GIS) as “the organization, presentation, communication and utilization of geo-information in graphic, digital or tactile form. It can include all stages from data preparation to end use in the creation of maps and related spatial information products”. In the context of map reproduction, it is important to note that this definition also takes into account the technological aspect in the cartographic process and highlights the importance of it in the creation of spatial information products beyond maps.

According to MacEachren [20], apart from their obvious role of visualising geodata, maps also serve as interfaces to the underlying computations, connect human reasoning to complex sources of information and facilitate the understanding of spatial relations. The current map making practices in academia use various software products for different tasks, from spreadsheets to specialized statistical software to specialised cartographic packages and geographic information systems. Frequently, the cartographic process is broken down into multiple steps [6], which are not always connected in a straightforward way. These steps can include data management, statistical analyses, geoprocessing and graphical display of the results. While there is a lot of variability involved – in particular in a scientific context, each step needs to be validated for its objectivity and ideally should be reproducible [7].

2.2 Open Science Practices

A fundamental requirement for enabling the reproduction of maps or computational workflows in general is access to all components necessary to carry out the reproduction (e.g., data, code, instructions) [23]. The central importance of the availability of research materials (especially data) is reflected in the several initiatives to standardise the way in which they are publicly shared. The FAIR Data Principles are guidelines for making data Findable, Accessible, Interoperable, and Reusable and have received considerable support by many scientific communities [36]. The Force11 community [2] also introduced the Principles for Data [9] and for Software Citation [33], acknowledging the significance of making these assets accessible. These principles are slowly being adopted by publishers as well. The publisher Copernicus Publications², for example, enforces the citation of data and encourages the citation of software, referencing the aforementioned principles.

The Transparency and Openness Promotion (TOP) Committee defined a set of standards with increasing levels of stringency to facilitate and motivate an open culture in scholarly communication [22]. These standards, namely the TOP Guidelines, deal with the topics of data, code and other materials citation and sharing, as well as the transparency of study design (among others). They indicate the extent to which a journal considers them as a requirement for publishing. Based on the TOP Guidelines, the Center for Open Science uses the TOP Factor³ rating to rank journals [31]. Journals from the disciplines of Geography, Planning and Development, and Earth and Planetary Sciences score seven at maximum in this scale (Cartography and Geographic Information Science, Nature Geoscience, Nature Sustainability), while the highest achieved score across all disciplines is 27.

2.3 Reproducible Research in the Geosciences

Researchers have started to address the issue of reproducibility in the Geosciences and have recognised that open data and methods and open source software are prerequisites for achieving the full potential in transparency [24]. However, Ledermann and Gartner [19] argue that acquiring the source code of an experiment is not enough to reproduce it and rather argue that a well defined and clearly structured programming workflow in an ontological fashion contributes to the transparency, reproducibility and extensibility of scientific experiments. Giraud and Lambert [6] encourage the use of literate programming reports, e.g., Jupyter notebooks or R Markdown, because such programming solutions provide complete instructions from raw data to the cartographic product. In a later work, the same

² <https://publications.copernicus.org/>

³ <https://topfactor.org/>

authors [7] provide an example of reproducible cartographic workflow, which is implemented by combining different geospatial R packages in a single R Markdown script. Knoth and Nüst [14] leverage containers for describing the computational environment of the experiment and for facilitating the reuse of their workflow by others. In more complex visualisation cases, such as terrain representation, Kennelly et al. [13] commented on the usefulness of 3D reference models for evaluating and comparing reproduced visualizations and stated that in the field of cartography such data models do not exist.

Reproducing and replicating visualizations in general is not an easy task. Fekete and Freire [5] consider interactive features as an additional complexity for scientific replication, but emphasize the importance of interactivity in data exploration. A data-centric approach for reproducible visualizations has been proposed by Silva et al. [32], where the authors highlight the significance of a well-defined data flow pipeline. Irrespective of the implementation tools and description mechanisms of the data flow, the purpose of the visualization process is always to gain insights from the data.

2.4 Reproducible Map Making

Even though reproducibility is a topic of rising concern for the scientific community, the term “reproducible cartography” is not frequently encountered in the literature. Giraud and Lambert [6] place reproducible maps on a spectrum, ranging from non-reproducible, when they come as a simple print-out, to fully reproducible, when they are clearly linked with executable code, data and metadata. Although many suggestions have been made regarding the enhancement of map reproducibility in different contexts (terrain representation [13], knowledge graphs [21], etc.), core terms such as “reproducible cartography” and “map reproducibility” so far have not been clearly specified, let alone formally defined.

Since the widely used ICA definition of cartography as a discipline also includes an artistic dimension [17], we prefer to use the terms “map production” and “map making” in the context of this paper and map reproducibility in general. Based on the definition of Barba [1] for reproducible research outlined in Section 1 and on the definition of Taylor [35] for cartography, we propose the following definition:

Reproducible map production/making refers to the provision, organisation, and processing of all materials used in the cartographic process in such a way that a map as a cartographic product can be recreated in an independent experiment.

This process thus aims to create a visual copy of the original map without introducing any significant variations that alter the maps’s interpretation. Although the term *map* can refer to a variety of representational artifacts, in the context of this work we focus solely on maps with an apparent geographical reference.

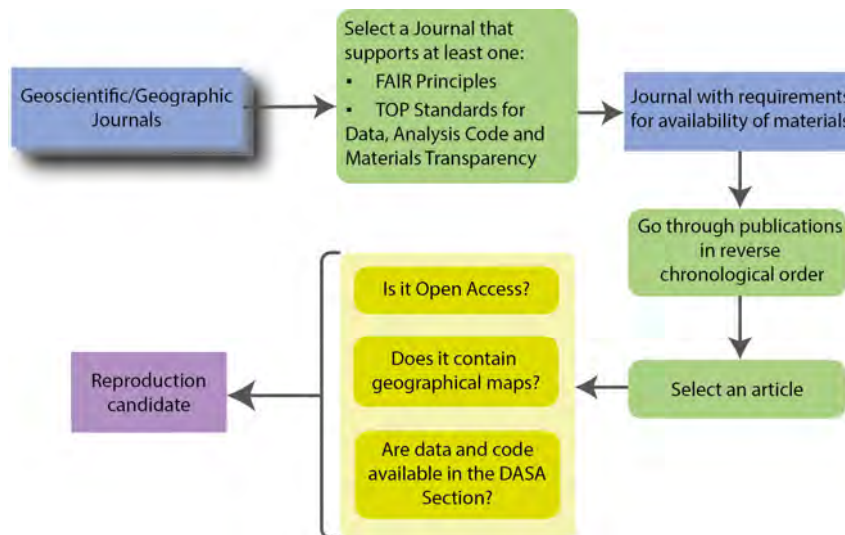
3 Methodological Approach

The main goal of our approach was to assess the degree to which map making practices in scientific publications allow for successful reproduction of the included maps. For this purpose, we collected a set of papers and subsequently attempted to reproduce their map figures. During this process, we screened which tools and programming languages were commonly used for map creation. We also collected information about the effort we had to invest during our reproduction attempts. All the reproductions were run on laptop running Linux Ubuntu 20.04 with 16 GB RAM and an i7-1185G7 @ 3.00GHz with 8 cores.

3.1 Creating the Paper Collection

Selecting a good sample of publications for our study was not an easy task due to the diversity of geoscientific papers and the maps used therein. In order to capture current practices from a broad range of journals, we limited our search to recent articles that were available via Open Access (OA) and were published within the 12 months prior to the beginning of our study (June 2022). The selection workflow that we eventually implemented is illustrated in Figure 1. Although maps can in principle be a part of any paper that reports on research in a geographical context, our investigation targeted journals with an explicit geographical or geoscientific scope. Aligning with the findings outlined in Subsection 2.2, we considered journals that support either the FAIR Data Principles or comply with the TOP Guidelines for Data, Code and Materials Transparency in order to maximise our chances of obtaining all the necessary components for map reproduction. In addition, eligible journals should require a data and/or software availability statement for publishing.

After selecting a journal, the research articles and data description papers were screened chronologically starting from the most recently published and working backwards until the cut-off date (June 2021). More recent works were preferred assuming that the related materials, namely code and data, are more probable to be available and the authors' contacts up-to-date when we ran the study. We excluded technical reports, briefings and reviews to ensure a consistent body of core scientific publications. In order for an article to qualify as suitable candidate for reproduction, we also required it to be OA so that anyone would be able to reproduce the illustrated maps. The next step was scanning the articles to confirm that geographical maps are included. Finally, the data and software availability statement had to clearly mention sources for acquiring the data and the code. If all the aforementioned conditions were met, we added this paper as a reproduction candidate to the list of papers we tried to reproduce.



■ **Figure 1** Flowchart of the paper selection process.

3.2 Reproduction Protocol

The procedure that we followed for the actual map reproduction is shown in Figure 2. The process started by reading and following the instructions in the availability statement to obtain the datasets and the software that were used in the paper analysis. This first stage

could involve several different steps. In the simplest scenario, the statement contained links where datasets and software could be directly downloaded. The data was found either in a repository maintained by the authors or in a public data portal such as the Copernicus Climate Data Store⁴. In the first case, the data was usually (pre)processed to some extent by the authors, while in the latter, the reproducing researcher had to perform all the data processing stages again. In some cases, the statement briefly explained why the datasets cannot be publicly shared, optionally encouraging the reader to get in contact with the authors or with a third person that is responsible for the distribution. The code was shared in the same repository as the data or in a different one, for example a public website specialising in hosting code such as Github⁵. The repositories occasionally contained directions for connecting the different components to setup the computational environment and re-run the map generation process. After collecting all the necessary materials, we identified the datasets and the parts of the code that corresponded to the creation of every map in the paper. This step required a deeper understanding of the paper and its maps. In particular, this involved reading the figure-related parts of the paper carefully to comprehend the map's purpose in the paper (e.g., exploratory data analysis, presenting results, placeholder map for GUI demonstration) and to determine its role in the computational pipeline.

As a next step, we then attempted to setup the computational environment by installing the required applications, libraries and plugins. If the attempt failed, we investigated the underlying issues and tried to solve them. Potential reasons for failure at this stage were missing data and unclear or incomplete instructions. When all the components were properly assembled and the computing pipeline was set up, we ran the analysis. We considered this step successful if it led to a map output. If running the analysis failed, we repeated the troubleshooting process. Possible reasons could be again missing data files, different library versions and different data processing results in preceding steps.

As a final step in the reproduction process, we assessed how well the generated map reproduced the map published in the original paper. For this purpose, we used a set of assessment criteria, which are described in Subsection 3.2.1. If a generated map was considered a successful reproduction according to these criteria, we added it to the list of successfully reproduced maps. Otherwise, we investigated the reasons for failure and attempted to address them until we ran out of options. If during the entire reproduction process information was unclear, we first tried to infer it from information in the paper and on from online sources, and then contacted the corresponding author to ask for their help. In general, we minimized the amount of intervention in the map production process and avoided adding data transformation steps or visualisation adjustments that were not explicitly stated by the authors of the paper.

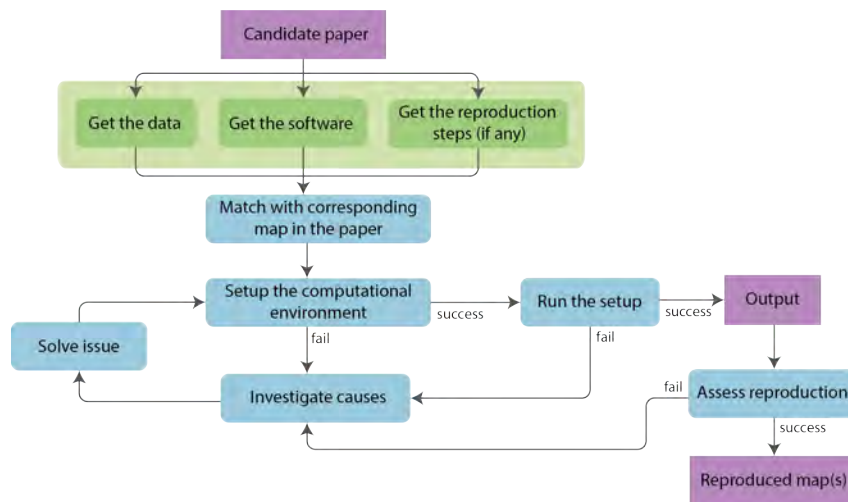
3.2.1 Assessing Reproduction Success

To confirm if a map reproduction was successful, initially we used the utility `compare` from the package `ImageMagick`⁶ to perform this image-based comparison, but the differences in image resolution, file types and file sizes led to quite large deltas between the original and the reproduced maps, even when they were visually and/or semantically very similar. This was partially due to the fact that the scripts frequently did not generate the same file types

⁴ <https://cds.climate.copernicus.eu/>

⁵ <https://github.org>

⁶ <https://imagemagick.org/>



■ **Figure 2** Flowchart of the reproduction process.

as the ones we found in the embedded figures on the publisher’s website. For this reason, we avoided the use of any tool for our final assessment and visually compared the reproduced figures with the original ones in a side-to-side fashion.

It is important to emphasise that even when two maps (the original and the reproduced one) are not exactly identical, they still can manage to convey the same message. Any criteria for assessing map reproduction should therefore go beyond simple pixel-based comparison. Based on these considerations, previous work on the reproducibility of figures in scientific publications [16] and the experiences gained during our reproduction study, we used a three by two matrix to assess the success of map reproduction (see Table 1). In this matrix, we categorised the observed differences based on the map element which they concern: the map body, the legend and other elements. The *map body* refers to the actual depiction of a geographic region in the map, e.g. a visual abstraction of a country. The *legend* contains complementary information that define the meaning of entities visualised in the map body. This includes, for example, a table explaining the meaning of symbols shown in the map body. All other visual elements, which are neither part of the map body nor the legend, were grouped under the *other elements* category. This includes, e.g., the North arrow, scale bars, or map titles.

For each of the three categories, we distinguish between two types of differences: aesthetic and semantic ones. *Aesthetic differences* refer to the way in which the map elements are visually expressed or styled, e.g., the colour used to depict certain aspects or which font was used. *Semantic differences* include any changes that affect the substance of the map, e.g., the absence of relevant elements or factual differences. When assessing whether a map reproduction was successful, we used the matrix to classify any differences we observed. If the observed differences were aesthetic in nature and were consistent with the context of the paper, we considered the reproduction a success. For example, a different colour scheme in the reproduced map is an aesthetic difference and, if reflected in the legend as well, it does not break the success of the reproduction. If not, though, it constitutes a semantic difference and the map reproduction does not qualify as successful. It should be mentioned that the success of map reproduction also depends on the purpose, the context and the target audience of the initial map.

■ **Table 1** Map elements and categories used to assess reproduction success with general examples (in italics) and examples from reproduction study (in bold); numbers in brackets indicate how often a difference type occurred during our reproduction study.

	Aesthetic	Semantic
Map body	colour (4), annotation style (3)	geographic extent (4), annotation placement (14), spatial distribution (2)
Legend	font (1), placement (1), <i>colour</i>	measurement scale (1), <i>text content, mismatch with map body</i>
Other elements	font (3), placement (10), stroke width (1), <i>size, colour</i>	<i>measurement units</i>

4 Results

The paper collection process described in Section 3 resulted in 27 reproduction candidates. The maps in two papers were demonstrating user interfaces and did not visualise any analysis results related to the paper content. They were thus not further considered. Ten papers on the list of reproduction candidates produced their maps using proprietary GIS or statistical software that was not freely distributed for educational or research purposes. In addition, some of these applications were unavailable for our (open source) operating system, i.e. Linux Ubuntu. This prevented us from reproducing the maps depicted in these papers in the way they were originally generated. However, three of these papers shared the scripts that they developed for their studies. One paper elaborated the data analysis and the relevant map creation using a proprietary GIS in detail, and shared these step-by-step instructions. This enabled us to attempt reproducing the map using a free and open source GIS. Six of the remaining 15 papers did not disclose any code at all, which also prevented us from reproducing the maps they contained.

4.1 Reproduction Outcomes

In total, we initiated the reproduction process as described in Figure 2 for nine out of the 27 reproduction candidates. Four of these reproductions were stalled after the phase of getting the data and the software. The reproduction of three candidate papers could not proceed because no part of the code was associated with creating the maps shown in the paper. For one candidate paper, reproduction was not possible because the code was compressed as a .rar file that produced a “corrupt header” error when unpacking. For the remaining six reproduction candidates, we were able to eventually generate a visual output. For two of these papers, we assessed the reproduction attempts of all the included maps as failed. We therefore managed to successfully reproduce maps from four papers.

The time required for reproduction varied greatly. One paper pointed to an online interactive notebook written in JavaScript that visualised the map figures of the paper in a transparent way. In this case, the successful reproduction required only a few clicks. For another paper, setting up the computational environment, going through several processing stages and resolving the issues we stumbled upon in cooperation with the authors, led to a successful reproduction only after seven weeks. The effort required for the other papers (both with failed and successful reproduction) fell between those two extremes. In the following, we describe the obstacles that we faced during the reproduction process in more detail.

4.2 Encountered Obstacles

In addition to issues preventing us from reproducing maps at all as outlined in 4.1, we mainly encountered data-related issues and issues in configuring the computational environment.

4.2.1 Data-related Issues

As discussed in Section 3, we targeted only papers with available data. Despite this, we still occasionally ran into missing files while executing the workflow. Since we could overcome such obstacles only with the help of the authors, we approached them via e-mail asking for information about where to find the missing files.

As mentioned before, some papers pointed to organisation websites for downloading the data. Possible changes in the organisation's data cleaning and pre-processing practices since the elaboration of the study led to different datasets, and eventually failed reproduction. The same was true for data processing methods that contain stochastic operations, which were neither preserved nor documented. Since substantial differences in the underlying datasets entail different analysis results and consequently maps, we stopped our efforts in such cases and considered the map reproduction for the corresponding paper as failed.

In the case of one paper, the data was shared as a database dump. Importing it into a database turned out to be complicated, as there was no accompanying information regarding the database name and log-in credentials. We eventually discovered the necessary information in the code files. A further issue was the lack of documentation of the installed plugins in the database where the dump was extracted from. To solve this issue, we searched the Internet for the error/warning log we received when trying to work with the database dump to identify the missing plugins and to import the dump into a properly configured database.

4.2.2 Computational Environment Configurations

While data-related issues caused problems for our reproduction attempts, by far the most considerable impediment that we faced in almost every reproduction attempt was related to the configuration of the computational environment. This includes in particular the lack of documentation regarding a) the versions of the software packages that make up the computational environment, b) the intended usage of the code, and c) the connections between the scripts and the data. This was especially true when there were multiple data processing steps involved in the map making process and several data sources that needed to be accessed individually and then combined. In order to resolve these issues, we had to resort to thoroughly scrutinising the source code itself and to experimenting with various configurations in the hope to identify the correct one. Frequently, we also had to change the (hard coded) file paths in the code in order to match our file system structure.

Such issues could be addressed by containerisation, for example, but we did not come across any paper that preserved the computational environment – neither in the form of a container nor as a deployed application. One of the screened papers shared the package list for its Python environment as .yml file, which saved us much time. However, we had to manually change the version of one package, since there were conflicts that could not be automatically resolved by the package manager (conda). Two papers listed the names of the Python packages they considered most significant, but without mentioning the corresponding versions. This caused substantial extra effort during the reproduction process as some of the packages we had to install in our system conflicted with each other during the execution of the scripts, resulting in an abrupt termination of the workflow. After an extensive Internet research on various Python forums, we tracked down which packages caused the error and tried several combinations of different versions until we could execute the scripts successfully.

Although the aforementioned issues mainly refer to interpreted languages that rely on package availability, such as Python and R, similar problems were encountered in the case of applications that use compiled languages, such as Java. Changes in the version or linkage to the repositories of the application's dependencies (libraries, plugins) not only mandate re-configuration in part of the code, but also may result in different visualised output.

4.3 Communication with the Authors

Communicating with the authors proved to be a valuable resource in addressing many of the issues outlined in Subsection 4.2. We contacted the correspondence authors via the e-mail address provided in the paper, stating that we are running a reproduction study that is focused on maps and expressing our interest in their work. In the same e-mail we described the technical issues that we were facing and asked for their help. Most frequently we reached out for help in case the code required more data files than those pointed out by the availability section. When we came across broken links, we requested the correct URLs from the authors. Other reasons for contacting the authors were to obtain more detailed instructions for executing the code and to clarify which library versions had to be used. It is worth noting that the authors almost always replied to us within a few days and were willing to advise on how to proceed with the reproduction. In one occasion, this communication resulted in an update of the corresponding data repository of the paper. We are very grateful for the extensive and helpful support we received from the authors during this study.

4.4 Reproduced maps

Successful Reproduction

The maps in Figure 3 were reproduced using R scripts [12]. We can observe that the reproduced map has titles for all the subfigures, while the original does not. The fonts of the axes labels and units and the legend are not the same. The units in the axes that show the longitude and latitude in degrees are represented in a different way. The orientation of the vertical axis labels is also different. Finally, the strokes appear bolder and the colors more saturated in the reproduced map. As we can see, the original and the reproduced maps are not identical in the pixel level, but they visualise the same data in a way that the intended message is still conveyed within the context of the paper. Therefore we considered this reproduction to be successful.

Unsuccessful Reproduction

In this example, the analysis of the paper was elaborated in ArcGIS Pro. Taking advantage of the instructions for reproducing the analysis that was provided by the authors [29], we tried to reproduce these steps in QGIS. Evidently, the reproduced maps differ greatly from the original ones. We attribute these deviations to different implementations in the underlying functionalities of the two GUI applications and to possible missing steps in the instruction sheet. Apart from the color scheme, the fonts, the basemap and the geographic extent, which we attempted to approximate with manual configurations in the GUI, we ended up with deep semantic differences that change drastically how the maps are read in the context of the paper. The reproduced maps differ in the shape of the tiled area, the measurement scale (as observed in the legend), and eventually, in the spatial distribution that is illustrated. As the maps diverge considerably in semantic level, we considered this reproduction attempt unsuccessful.

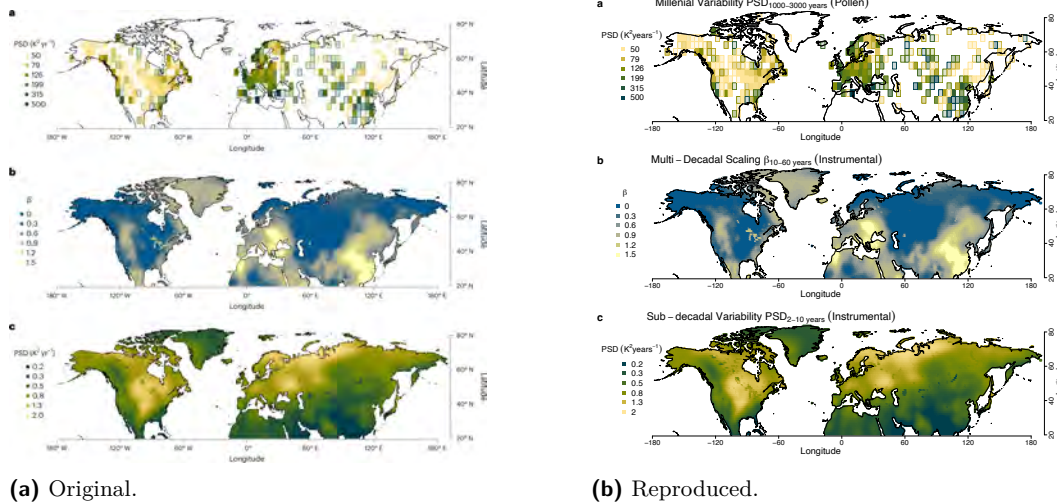


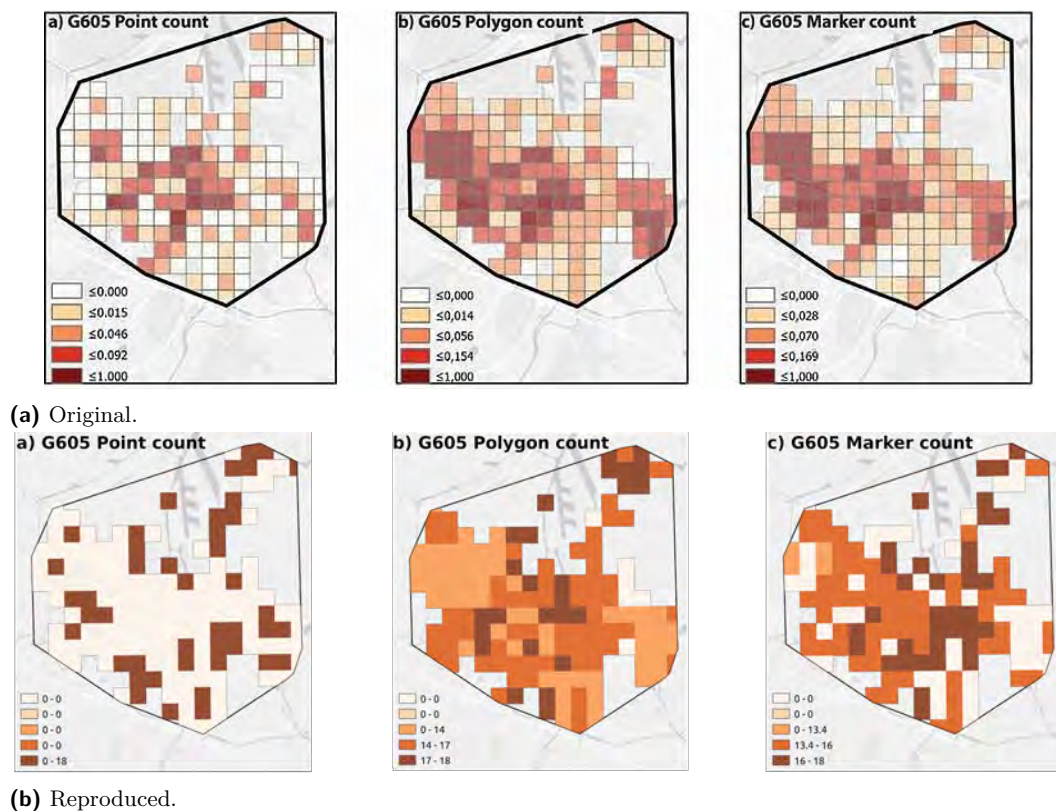
Figure 3 Example of successful reproduction with differences of aesthetic nature. Original (a) and reproduced (b) maps were created with R script. Original Figure (a) is extracted from Herbert et al. [11] as is, under CC BY 4.0 (<https://creativecommons.org/licenses/by/4.0/>).

5 Discussion

5.1 The State of Map Reproducibility

Based on our results and the insights gained through the reproduction study, it is evident that map reproducibility is in a dire state: out of 27 candidate papers from the last two years, we were able to only reproduce maps from four papers. However, it should be noted that reproducibility in general is not yet fully established in the Geosciences as a general requirement. In addition, half of the papers used commercial, proprietary software, which we excluded since it would require those who want to reproduce the contained maps to purchase the software. It is possible that the maps of those papers could have been reproduced, had we accepted to use the corresponding software. In one such case, we attempted to reproduce the maps using free and open software, which was not successful. Standardising map production functionalities and its description could be a way to bridge this gap, as it would enable different software to generate the same maps. Overall, we were positively surprised to be able to reproduce any maps at all, given the obstacles we faced.

Conceptualising and operationalising map reproduction success was also a challenge, as it became clear quite quickly that a pixel-based comparison is not rather useful. The initial set of criteria we defined enabled us to systematically assess reproduction success but constitute only a first step towards defining and formalising this concept and the underlying process. On a more practical level, poor documentation or a complete lack thereof required the most effort and frequently led to reproduction failure. While this is understandable given the current reward mechanisms in scientific publishing, it would also be easy to overcome and greatly reduce the effort involved in map reproduction. The data related issues we encountered also highlighted the importance of archival data repositories with persistent links for map reproduction.



■ **Figure 4** Example of unsuccessful reproduction with semantic differences. Original (a) was created with ArcGIS Pro and shows a subset of Figure 5 of Ramírez Aranda et al. [30], under CC BY 4.0 (<https://creativecommons.org/licenses/by/4.0/>). The reproduction (b) was created with QGIS.

5.2 Recommendations for Improving Map Reproducibility

From the experiences we gathered during our reproduction, we can infer a number of recommendations for improving map reproducibility. The most important is to *fully document all steps and intermediate results*, so that people unfamiliar with the research can execute them independently and assess whether the intermediate outcomes they produce correspond to those produced by the authors, in line with what was proposed by Silva et al. [32]. In the simplest scenario, this could include providing a flowchart that shows which data was used at what step using which analysis component. This way, it is much easier to identify what might have gone wrong when the final result differs from the original map. A second recommendation for improving map reproducibility is to *publish data, code and configurations with reproduction in mind*. Using persistent links and data repositories avoids issues related to dead links and changed datasets. When using public repositories, researchers should assess their suitability for reproduction and make use of features that facilitate reproduction (e.g., archiving repositories in Github so that the snapshot at map production time is preserved). Furthermore, it is helpful to provide relative file paths rather than absolute ones as well as to preserve the structure of the file system. Containerisation is one promising option that can help overcome many issues related to computational environment configurations, which has been successfully used for reproduction in previous work [24]. A third recommendation is to *keep the map production process as simple as possible*. When we tried to reproduce maps, we

generally observed that the more steps were involved (from data pre-processing to the final map), the more difficult it became to successfully reproduce the final map. In some cases, complex production processes cannot be avoided but should then be very well documented. A fourth recommendation is to *educate researchers on how to conduct reproducible research*. This has been suggested before [16, 26] and is important not only for making them aware of the importance of reproducibility, but also to enable them to publish their research in a reproducible way, including the maps they develop. Our final recommendation is to *define reproducible map standards* so that assessing successful reproduction becomes easier and researchers can easily estimate whether their published work meets those standards. There are suggestions for reproducibility standards in general [22, 23, 27], but as outlined in Section 3.2.1, a more nuanced approach might be needed for maps as a simple pixel-based comparison does not capture well whether two maps convey the same message. The criteria we outlined in that section are a starting point towards defining such standards. Finally, a change of reward schemes in academic publishing (also strongly recommended by Ostermann et al. in [27]) could greatly benefit map reproducibility, as at the moment the extra effort required for making maps reproducible is not rewarded.

5.3 Limitations

Our study was exploratory in nature and thus is subject to a number of limitations. This includes the relatively small amount of papers (27), which was reduced further after deeper analysis. While this limits the generalisability of our findings, we were still able to identify many relevant obstacles and gain insights into how maps from academic papers can be reproduced. In addition, the reproduction study was carried out without a rigid, predefined protocol, which was due to the novel domain (map reproducibility). Hence, later reproduction attempts benefited from lessons learnt in earlier ones. Furthermore, the study was carried out by a single researcher and their abilities and knowledge affected the reproduction process, e.g., in terms of expertise in certain programming languages. It can be expected that reproducing researchers vary in their abilities and knowledge as well, so while having only one person do the reproductions limited generalisability, it also provided a realistic test for map reproducibility. Finally, we used an initial simple set of assessment criteria to determine the success of map reproduction. While clearly further research is needed here, we consider the proposed set a good starting point and the presented study can serve as an initial evaluation of those criteria as well.

5.4 Future Work

Our work brought to light several gray areas and loosely defined concepts regarding reproducible research, which also depend on individual interpretation. An important area for future work is thus to further investigate the diverse perspectives surrounding the notion of successful map reproduction. This applies in particular to the question when a map is considered reproducible, which factors affect this decision, and how significant various semantic and aesthetic differences are in this context. Another interesting question for further research is to what extent can the reproducing researcher modify the map making process to create an identical copy of the original map without compromising its integrity. Understanding and integrating the different viewpoints will contribute to a more comprehensive evaluation of map reproducibility. The clarification of these concepts will pave the way for more systematic approaches to map comparisons, which can be useful beyond the reproducibility of the map making processes as well. Examining these factors will contribute to advancing the understanding and implementation of reproducible research in the Geosciences.

6 Conclusion

In this study, we explored how reproducible maps are that are included in recent scientific publications. We collected a total of 27 papers, attempted to reproduce the maps contained therein, and managed to successfully reproduce them in four cases. We report on the process that we followed and the obstacles that we encountered. Our key contributions are an initial definition of *reproducible map making*, an inceptive set of criteria to assess the success of a map reproduction attempt and a set of guidelines for improving map reproducibility in geoscientific publications. Our work – while exploratory in nature – provides a first systematic analysis of map reproducibility. These outcomes are aligned with previous work regarding reproducibility in the domain of Geosciences but the particularities of maps as representational artifacts introduce further challenges and require further research for conceptualising and operationalising map reproduction. As a next step in this line of work, we are therefore planning to consult researchers in the Geosciences to develop a deeper understanding of what constitutes a successful map reproduction.

References

- 1 Lorena A. Barba. Terminologies for Reproducible Research, February 2018. doi:10.48550/arXiv.1802.03311.
- 2 Philip E. Bourne, Timothy W. Clark, Robert Dale, Anita de Waard, Ivan Herman, Eduard H. Hovy, and David Shotton. Improving The Future of Research Communications and e-Scholarship (Dagstuhl Perspectives Workshop 11331). *Dagstuhl Manifestos*, 1(1):41–60, 2012. Place: Dagstuhl, Germany Publisher: Schloss Dagstuhl — Leibniz-Zentrum fuer Informatik. doi:10.4230/DagMan.1.1.41.
- 3 Jon F. Claerbout and Martin Karrenbach. Electronic documents give reproducible research a new meaning. In *SEG Technical Program Expanded Abstracts 1992*, SEG Technical Program Expanded Abstracts, pages 601–604. Society of Exploration Geophysicists, January 1992. doi:10.1190/1.1822162.
- 4 David L. Donoho, Arian Maleki, Inam Ur Rahman, Morteza Shahram, and Victoria Stodden. Reproducible Research in Computational Harmonic Analysis. *Computing in Science & Engineering*, 11(1):8–18, January 2009. Conference Name: Computing in Science & Engineering. doi:10.1109/MCSE.2009.15.
- 5 Jean-Daniel Fekete and Juliana Freire. Exploring Reproducibility in Visualization. *IEEE Computer Graphics and Applications*, 40(5):108–119, September 2020. doi:10.1109/MCG.2020.3006412.
- 6 Timothée Giraud and Nicolas Lambert. Reproducible Cartography. In Michael P. Peterson, editor, *Advances in Cartography and GIScience*, Lecture Notes in Geoinformation and Cartography, pages 173–183, Cham, 2017. Springer International Publishing. doi:10.1007/978-3-319-57336-6_13.
- 7 Timothée Giraud and Nicolas Lambert. Reproducible Workflow for Cartography – Migrants Deaths in the Mediterranean. *Proceedings of the ICA*, 2:1–7, July 2019. doi:10.5194/ica-proc-2-38-2019.
- 8 Michael F. Goodchild, A. Stewart Fotheringham, Peter Kedron, and Wenwen Li. Introduction: Forum on Reproducibility and Replicability in Geography. *Annals of the American Association of Geographers*, 111(5):1271–1274, July 2021. doi:10.1080/24694452.2020.1806030.
- 9 Data Citation Synthesis Group. Joint Declaration of Data Citation Principles. Technical report, Force11, 2014. doi:10.25490/A97F-EGYK.
- 10 Michael A. Heroux, Lorena Barba, Manish Parashar, Victoria Stodden, and Michela Taufer. Toward a Compatible Reproducibility Taxonomy for Computational and Computing Sciences. Technical Report SAND2018-11186, Sandia National Lab. (SNL-NM), Albuquerque, NM (United States), October 2018. doi:10.2172/1481626.

- 11 R. Hébert, U. Herzschuh, and T. Laepple. Millennial-scale climate variability over land overprinted by ocean temperature fluctuations. *Nature Geoscience*, 15(11):899–905, November 2022. Number: 11 Publisher: Nature Publishing Group. doi:10.1038/s41561-022-01056-4.
- 12 Raphaël Hébert. Hhl2022, September 2022. doi:10.5281/zenodo.7062762.
- 13 Patrick J. Kennelly, Tom Patterson, Bernhard Jenny, Daniel P. Huffman, Brooke E. Marston, Sarah Bell, and Alexander M. Tait. Elevation models for reproducible evaluation of terrain representation. *Cartography and Geographic Information Science*, 48(1):63–77, January 2021. doi:10.1080/15230406.2020.1830856.
- 14 Christian Knoth and Daniel Nüst. Reproducibility and Practical Adoption of GEOBIA with Open-Source Software in Docker Containers. *Remote Sensing*, 9(3):290, March 2017. Number: 3 Publisher: Multidisciplinary Digital Publishing Institute. doi:10.3390/rs9030290.
- 15 Markus Konkol and Christian Kray. In-depth examination of spatiotemporal figures in open reproducible research. *Cartography and Geographic Information Science*, 46(5):412–427, September 2019. doi:10.1080/15230406.2018.1512421.
- 16 Markus Konkol, Christian Kray, and Max Pfeiffer. Computational reproducibility in geoscientific papers: Insights from a series of studies with geoscientists and a reproduction study. *International Journal of Geographical Information Science*, 33(2):408–429, February 2019. doi:10.1080/13658816.2018.1508687.
- 17 Menno-Jan Kraak. Strategic Plan for 2019–2027. Technical report, International Cartographic Association, Enschede, April 2019. URL: <https://icaci.org/strategic-plan/>.
- 18 Menno-Jan Kraak and Sara Irina Fabrikant. Of maps, cartography and the geography of the International Cartographic Association. *International Journal of Cartography*, 3(sup1):9–31, October 2017. doi:10.1080/23729333.2017.1288535.
- 19 Florian Ledermann and Georg Gartner. Towards Conducting Reproducible Distributed Experiments in the Geosciences. *AGILE: GIScience Series*, 2:1–7, June 2021. doi:10.5194/agile-giss-2-33-2021.
- 20 Alan M MacEachren. Cartography as an Academic Field: A Lost Opportunity or a New Beginning? *Cartographic Journal*, 50(2):166–170, May 2013. doi:10.1179/0008704113Z.00000000083.
- 21 Gengchen Mai, Weiming Huang, Ling Cai, Rui Zhu, and Ni Lao. Narrative Cartography with Knowledge Graphs. *Journal of Geovisualization and Spatial Analysis*, 6(1):4, February 2022. doi:10.1007/s41651-021-00097-4.
- 22 B. A. Nosek et al. Promoting an open research culture. *Science*, 348(6242):1422–1425, June 2015. Publisher: American Association for the Advancement of Science. doi:10.1126/science.aab2374.
- 23 Daniel Nüst, Carlos Granell, Barbara Hofer, Markus Konkol, Frank O. Ostermann, Rusne Sileryte, and Valentina Cerutti. Reproducible research and GIScience: an evaluation using AGILE conference papers. *PeerJ*, 6:e5072, July 2018. Publisher: PeerJ Inc. doi:10.7717/peerj.5072.
- 24 Daniel Nüst, Markus Konkol, Edzer Pebesma, Christian Kray, Marc Schutzzeichel, Holger Przibytzin, and Jörg Lorenz. Opening the Publication Process with Executable Research Compendia. *D-Lib Magazine*, 23, January 2017. doi:10.1045/january2017-nuest.
- 25 Daniel Nüst and Edzer Pebesma. Practical Reproducibility in Geography and Geosciences. *Annals of the American Association of Geographers*, 111(5):1300–1310, July 2021. doi:10.1080/24694452.2020.1806028.
- 26 Frank O. Ostermann. Peer assessment to improve reproducibility of computational project work. In Hans-Ulrich Heiß, Hannu-Matti Järvinen, Annette Meyer, and Alexandra Schulz, editors, *SEFI 49th Annual Conference*, SEFI Proceedings, pages 1080–1090, Berlin, Germany, 2021. Technische Universität Berlin. URL: <https://www.sefi.be/wp-content/uploads/2021/12/SEFI49th-Proceedings-final.pdf>.

- 27 Frank O. Ostermann, Daniel Nüst, Carlos Granell, Barbara Hofer, and Markus Konkol. Reproducible Research and GIScience: An Evaluation Using GIScience Conference Papers. In Krzysztof Janowicz and Judith A. Versteegen, editors, *11th International Conference on Geographic Information Science (GIScience 2021) – Part II*, volume 208 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 2:1–2:16, Dagstuhl, Germany, 2021. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. doi:10.4230/LIPIcs.GIScience.2021.II.2.
- 28 Roger D. Peng. Reproducible Research in Computational Science. *Science*, 334(6060):1226–1227, December 2011. Publisher: American Association for the Advancement of Science. doi:10.1126/science.1213847.
- 29 Nohemí Ramirez Aranda, Jeroen De Waegemaeker, Viktor Venhorst, Wim Leendertse, Eva Kerselaers, and Nico Van de Weghe. Point, polygon, or marker? In search of the best geographic entity for mapping Cultural Ecosystem Services using the online PPGIS tool, “My Green Place.”, December 2020. doi:10.5281/zenodo.4347404.
- 30 Nohemí Ramírez Aranda, Jeroen De Waegemaeker, Viktor Venhorst, Wim Leendertse, Eva Kerselaers, and Nico Van de Weghe. Point, polygon, or marker? In search of the best geographic entity for mapping cultural ecosystem services using the online public participation geographic information systems tool, “My Green Place”. *Cartography and Geographic Information Science*, 48(6):491–511, November 2021. doi:10.1080/15230406.2021.1949392.
- 31 Center for Open Science. New Measure Rates Quality of Research Journals’ Policies to Promote Transparency and Reproducibility. URL: <https://www.cos.io/about/news/new-measure-rates-quality-research-journals-policies-promote-transparency-and-reproducibility>.
- 32 Claudio T. Silva, Juliana Freire, and Steven P. Callahan. Provenance for Visualizations: Reproducibility and Beyond. *Computing in Science & Engineering*, 9(5):82–89, September 2007. doi:10.1109/MCSE.2007.106.
- 33 Arfon M. Smith, Daniel S. Katz, and Kyle E. Niemeyer. Software citation principles. *PeerJ Computer Science*, 2:e86, September 2016. Publisher: PeerJ Inc. doi:10.7717/peerj-cs.86.
- 34 Daniel Sui and Peter Kedron. Reproducibility and Replicability in the Context of the Contested Identities of Geography. *Annals of the American Association of Geographers*, 111(5):1275–1283, July 2021. doi:10.1080/24694452.2020.1806024.
- 35 D. R. FRASER Taylor. CHAPTER 1 – Geographic Information Systems: The Microcomputer and Modern Cartography. In Fraser Taylor, editor, *Modern Cartography Series*, volume 1 of *Geographic Information Systems*, pages 1–20. Academic Press, January 1991. doi:10.1016/B978-0-08-040277-2.50009-X.
- 36 Mark D. Wilkinson et al. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(1):160018, March 2016. Number: 1 Publisher: Nature Publishing Group. doi:10.1038/sdata.2016.18.

Semi-Supervised Learning from Street-View Images and OpenStreetMap for Automatic Building Height Estimation

Hao Li ✉

Technical University of Munich, Germany

Zhendong Yuan ✉

Utrecht University, The Netherlands

Gabriel Dax ✉

Technical University of Munich, Germany

Gefei Kong ✉

Norwegian University of Science and Technology, Trondheim, Norway

Hongchao Fan ✉

Norwegian University of Science and Technology, Trondheim, Norway

Alexander Zipf ✉

GIScience Chair, Heidelberg University, Germany

Martin Werner ✉

Technical University of Munich, Germany

Abstract

Accurate building height estimation is key to the automatic derivation of 3D city models from emerging big geospatial data, including Volunteered Geographical Information (VGI). However, an automatic solution for large-scale building height estimation based on low-cost VGI data is currently missing. The fast development of VGI data platforms, especially OpenStreetMap (OSM) and crowdsourced street-view images (SVI), offers a stimulating opportunity to fill this research gap. In this work, we propose a semi-supervised learning (SSL) method of automatically estimating building height from Mapillary SVI and OSM data to generate low-cost and open-source 3D city modeling in LoD1. The proposed method consists of three parts: first, we propose an SSL schema with the option of setting a different ratio of “pseudo label” during the supervised regression; second, we extract multi-level morphometric features from OSM data (i.e., buildings and streets) for the purpose of inferring building height; last, we design a building floor estimation workflow with a pre-trained facade object detection network to generate “pseudo label” from SVI and assign it to the corresponding OSM building footprint. In a case study, we validate the proposed SSL method in the city of Heidelberg, Germany and evaluate the model performance against the reference data of building heights. Based on three different regression models, namely Random Forest (RF), Support Vector Machine (SVM), and Convolutional Neural Network (CNN), the SSL method leads to a clear performance boosting in estimating building heights with a Mean Absolute Error (MAE) around 2.1 meters, which is competitive to state-of-the-art approaches. The preliminary result is promising and motivates our future work in scaling up the proposed method based on low-cost VGI data, with possibilities in even regions and areas with diverse data quality and availability.

2012 ACM Subject Classification Information systems; Information systems → Geographic information systems

Keywords and phrases OpenStreetMap, Street-view Images, VGI, GeoAI, 3D city model, Facade parsing

Digital Object Identifier 10.4230/LIPICs.GIScience.2023.7

Supplementary Material

Software (Data and code supporting this paper): https://github.com/bobleegogogo/building_height; archived at [swh:1:dir:1731a2bf38d083320ed151eefd51b4c6686c3f7c](https://swh.io/dir/1731a2bf38d083320ed151eefd51b4c6686c3f7c)



© Hao Li, Zhendong Yuan, Gabriel Dax, Gefei Kong, Hongchao Fan, Alexander Zipf, and Martin Werner;

licensed under Creative Commons License CC-BY 4.0

12th International Conference on Geographic Information Science (GIScience 2023).

Editors: Roger Beecham, Jed A. Long, Dianna Smith, Qunshan Zhao, and Sarah Wise; Article No. 7; pp. 7:1–7:15

Leibniz International Proceedings in Informatics



Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Introduction

For decades, the world has been comprehensively mapped in 2D, however a vertical dimension remains underexplored despite its huge potential, which is even more critical in Global South areas due to inherent mapping inequality and diverse data availability. Mapping human settlements as a 3D representation of reality requires an accurate description of vertical dimension besides the 2D footprints and shapes [19, 11, 14, 7, 23]. Such 3D representation of human settlements is of significant importance in many aspects, for instance, quiet and shadow routing [35], environmental exposure modeling [2, 40, 34], architecture and city planning [32, 36] and population capacity estimation [37, 21]. However, it remains challenging to derive low-cost and open-source 3D representation of buildings at scale. In this paper, with “low-cost”, we mainly refer to the cost of data acquisition in 3D building modeling.

Given existing methods of photogrammetry and remote sensing, 3D city reconstruction is still a high-cost and time-consuming task, which mostly requires extensive expert knowledge and a large amount of geospatial data (e.g., cadastral data, airborne photogrammetry data). This fact will certainly increase the difficulty of ordinary stakeholders and city governments with limited funding in establishing 3D city modeling systems for their well-being demands. Fortunately, the increasing availability of Volunteer Geographic Information (VGI) together with crowdsourcing technology [16] has provided a low-cost and scalable solution of mapping our world even in a 3D representation. OpenStreetMap (OSM), as the most successful VGI project, was considered as a valuable global data source for creating large-scale 3D city models [14, 12]. For instance, in [10], a joint processing method of OSM and mutli-sensor remote sensing data (e.g., TanDEM-X and Sentinel-2) was developed to generate large-scale 3D urban reconstruction; Milojevic-Dupont et al [27]. demonstrated the capability of accurate building height prediction purely based on morphometric features (or urban forms) extracted from OSM data (e.g., building and street geometry).

Moreover, several recent works in [41] and [28] highlight the huge potential of low-cost street-view images (SVI) in increasing the efficiency of large-scale 3D city modeling. The idea is intuitive as SVI provides a low-cost and close-range observation of urban buildings, therefore contains key information needed for 3D reconstruction, such as facade elements, shapes, and building heights. Given the fast development of geospatial machine learning and artificial intelligence (GeoAI) [17], automatic interpretations of SVI have become more efficient than ever before. Hence, the geospatial ML method, which can integrate building height information derived from SVI with existing 2D building footprints from OSM, presents a promising solution for creating large-scale and open-source 3D city models.



■ **Figure 1** An overview of building height estimation via semi-supervised learning from OpenStreetMap data and street-view images.

In this paper, we propose a semi-supervised learning (SSL) method (as shown in Figure 1) to accurately estimate building height based on open-source SVI and OSM data. As a case study, we implement the proposed method by training three different machine learning (ML) models, namely Random Forests (RF), Support Vector Machine (SVM), and Convolutional Neural Network (CNN), in the city of Heidelberg, Germany. Specifically, we first extract multi-level urban morphometric features from existing OSM data (i.e., buildings, streets, street blocks) as a feature space to the regression of building height, then we collect SVI with metadata via the Mapillary platform (<https://www.mapillary.com>) and design a building floor estimation workflow with a pre-trained facade object detection network to generate “pseudo label” for the SSL of building height estimation models. As a result, we create an open-source LoD1 3D city models for selected areas in Heidelberg using the low-cost SVI data and OSM 2D building footprints.

2 Related Work

2.1 Building Height Estimation

Existing methods of building height estimation generally rely on Light Detection and Ranging (LiDAR) [15, 29], Synthetic Aperture Radar (SAR) [25], and high-resolution remote sensing image data [26]. In these data sources, LiDAR data provides highly accurate information of building height but is difficult to estimate building height in large scale, considering its collection cost. For SAR, the estimation result is often affected by the mixture of different microwave scattering, thus have high uncertainties [33]. To avoid these problems, many researchers also investigate remote sensing image data. For these methods, considering that remote sensing image data does not contain 3D information directly, existing works select stereo/multi-view images as the data source to achieve the estimation of building height [1, 8, 42].

However, although SAR and remote sensing image data have a relatively low collection cost than LiDAR data, the complex data processing of these data source causes their high time and labor costs. Compared with these three data, SVI data and 2D building footprint data are easier and cheaper to be collected and processed, especially with the support of VGI (e.g., Mapillary and OpenStreetMap). There have been some early efforts to estimate building height based on these new data sources. Biljecki et al. [6], Milojevic-Dupont et al. [27], and Bernard et al. [4] proposed several methods based on RF or other ML approaches to analyze the relationship between building heights and their features (such as building area and type), and finally achieve the building height estimation from 2D footprint data. Yan and Huang [39] proposed a deep learning-based method to estimate building height from SVI. Zhao et al. [43] combined 2D building footprints and SVI to estimate building heights, which also used deep learning technology. These methods also achieved good performance but require a large amount of training data, which limits their generalization and practicality. Currently, there is little work on how to accurately estimate building height from 2D building footprint and SVI with only limited training data.

2.2 VGI and 3D Building Models

CityGML is a well-known international standard for 3D building modeling. In CityGML 2.0, 3D building models are divided into five levels of detail (LoD). In LoD0, only the 2D footprint information is involved in the model. In LoD1, the LoD0 model is extruded by their building heights, and the obtained cuboid after extrusion are the LoD1 model. In

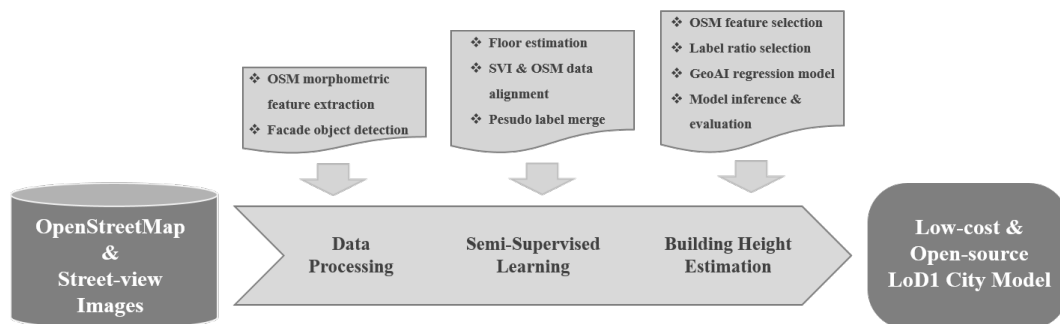
7:4 Automatic Building Height Estimation

LoD2, the 3D roof structure information is added into the LoD2 model. The LoD3 model further contains the facade element information, such as windows and doors. The LoD4 model is more complicated and contains both external and internal building elements. To meet the requirements of the abovementioned CityGML standard, many cities like New York, Singapore, and Berlin have created and freely released 3D city models with different LoDs in the past years. However, most of these 3D city building models are constructed in LoD1 or LoD2 for urban area, while large-scale and fine-grained (LoD3 and LoD4) models with semantic information are hardly available for cities with limited funding in establishing their own 3D city modelling systems. Hence, that is the main motivation of this work to provide a low-cost and open-source solution of creating large-scale 3D city models (e.g., first in LoD1).

Early work in [14] highlighted that OSM, as a crowdsourced VGI data source, can be combined with international standards of the Open Geospatial Consortium (OGC) to effectively create CityGML models in LoD1 and LoD2. Recently, Zhang et. al [41] proposed a web-based interactive system, namely VGI3D, as a collaborative platform to collect 3D building models with fine-grained semantic information in a crowdsourcing approach. In this work, we aim to further investigate the potential of low-cost VGI data sources, especially OSM data and crowdsourced SVI, in generating LoD1 3D city models via automatic building height estimation with only limited training data.

3 Methodology

The proposed method of automatic building height estimation mainly consists of three parts: (1) an SSL schema for height regression, (2) OSM morphometric feature extraction, and (3) building floor estimation based on the SVI. Figure 2 shows the methodological workflow of automatically generating open-source 3D city modeling (i.e., LoD1 city model) via the proposed SSL method. In the rest of this section, we will elaborate on the details of this design.



■ **Figure 2** The methodological workflow of automatic building height estimation from OpenStreetMap data and street-view images.

3.1 Semi-supervised Learning Schema

In traditional supervised learning, one relies on labelled data to build the prediction model. However, such a labelling process is mostly time consuming, labour demanding, and difficult to scale up. Therefore, the capability of learning from unlabeled data is a desirable feature to overcome this challenge. In this context, Semi-supervised learning (SSL) is a promising technique to accommodate the lack of labeled data by allowing the model to integrate part of

unlabeled data during the supervised model training [44, 18]. To be noticed, the SSL herein is different from self-supervised learning, which does not rely on any ground truth labels during the training process. A common way of implementing SSL is to generate “pseudo label” from the data itself or even auxiliary data [22], which can be then merged with existing labelled data to boost model performance. Following this concept, we design an SSL schema with the option of defining different ratio of “pseudo label” during the supervised regression of building height.

The proposed SSL schema is tasked with estimating building heights (h) based on a list of morphometric features $x = \langle x_1, \dots, x_m \rangle$ extracted from diverse scales of OSM data (e.g., individual building footprint, street network, street block, etc.), where m refers to the total number of features. In this context, the task of building height estimating can be formulated as a multifactor regression task in the following mathematic form:

$$h_{\Theta}(x) = \sum_{i=0}^m \Theta_i x_i \quad (1)$$

where $\Theta = \langle \Theta_1, \dots, \Theta_m \rangle$ is the corresponding regression coefficients. More importantly, the regression target value of building heights h comes from the following two parts:

$$h = (1 - a) * h_{Raw} + a * h_{SSL} \quad (2)$$

Where a is the ratio of “pseudo label” (h_{SSL}) obtained from automatic facade parsing of Mapillary SVI. We will elaborate on this later in Section 3.3, while it is sufficient to understand that besides available training label (i.e., known building heights) the model can also benefit from SSL labels which are extracted from large-scale and open-source SVI in an automatic and unsupervised method.

To build the model for accurate building height estimation, we train a classic supervised regression model of finding the optimal regression coefficients with gradient descent and optimizing a loss function of Mean Square Error (MSE) in the following format:

$$\mathcal{L}_{\Theta^*}^{MAE} = \arg \min_{\Theta} \frac{1}{N} \sum_{i=1}^N \| \hat{h}_i - h_i \| \quad (3)$$

where \mathcal{L}_{MAE} and Θ^* refer to the loss function and the optimal coefficients set, respectively, and N is the number of training samples (h_{SSL} and h_{Raw}).

The design of SSL is concise and model-independent, which means in case we can keep feeding the ML models with “pseudo label” (h_{SSL}) of building height extracted from SVI, the regression task can be tackled with diverse ML models. In this paper, we demonstrate the capability of these three ML models (i.e., RF, SVM, and CNN) in estimating building height in a typical western European city, so to say the city of Heidelberg, Germany.

3.2 OSM Morphometric Feature Extraction

Intensive existing works have confirmed the excellent capability of multi-level morphological features (or urban-form features) in predicting key attributes (e.g., height, function, energy consumption, etc.) of buildings and streets from an urban analytic perspective [27, 5].

To infer building height, we implement a range of morphometric features extracted from OSM at three different levels, namely building-level, street-level, and street block-level, as shown in Table 1. In total, we calculate 129 morphometric features based on OSM data (i.e., individual building footprints and street networks) to construct their spatial and geometric relationships (e.g., spatial vicinity and compactness of street-blocks). More specifically, we elaborate on the details of OSM morphometric features (in three distinct levels) as follows:

Building-level. Considering the hidden information from the building footprint itself, we calculate 9 features such as footprint area, perimeter, circular compactness, convexity, orientation and length of wall shared with other buildings. The intuition herein is that such building-level features can provide explicit and implicit information about the footprint shape (e.g., compactness and complexity), which contributes to estimating building heights. For instance, it was reported that a higher building generally consists of a large net internal area, and vice versa [6]. In addition, since buildings are mapped differently in OSM (e.g., one building in several polygons or several buildings in one polygon), we simplify this data quality issue by considering each polygon as a single building, while future work is definitely needed in investigating the impact of how individual buildings are presented in OSM.

Street-level. Besides morphometric features of the building footprint itself, the street network surrounding a building can be informative in estimating building height. For instance, a high density (or compactness) of streets can imply more high-story buildings in order to accommodate a potentially higher number of residents. Therefore, we calculate 9 features based on the spatial relationship of buildings and their closest streets and road intersections, such as length, average width, distance to the building, local closeness, betweenness and centrality, etc.

Street block-level. Furthermore, we generate morphological tessellations based on the OSM street network. This tessellation representation and its interaction with roads and buildings were included in the design of the feature space (8 features). The motivation is straightforward, as a preliminary assumption is that buildings in the same block are more likely to be of a similar height.

Moreover, to capture the spatial auto-correlation in the OSM data, we extend these three levels of OSM morphometric features by considering their second-order features (e.g., total, average, and standard deviation) in the neighbourhood (i.e., within 20, 50, and 500 meters buffers). As for the implementation, we rely on the open-source Python software toolkit called momepy v.0.5.1 to calculate these features. For a complete list of OSM morphometric features, please refer to the GitHub repository (https://github.com/bobleegogogo/building_height).

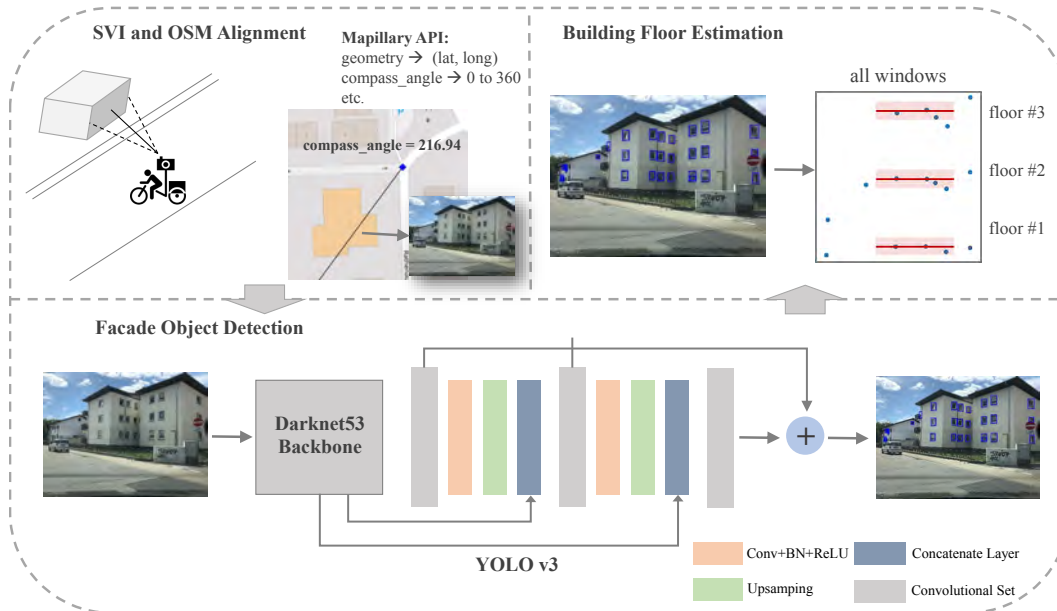
■ **Table 1** List of OSM morphometric features extracted at building-level, street-level, and street-block level.

Level	Group	Features	Count
Building	building footprint	e.g., area, perimeter, convexity etc.	9
	buildings within 50m	e.g., total/average/standard deviation of area, perimeter, convexity etc.	18
	buildings within 200m	e.g., total/average/standard deviation of perimeter, etc.	18
	buildings within 500m	e.g., total/average/standard deviation of convexity, etc.	18
Street	closest streets and intersections	e.g., length, closeness and distances to the intersection	9
	closest streets and intersections within 50m	e.g., total/average/standard deviation of distances to the closest intersection	11
	closest streets and intersections within 200m	e.g., total/average/standard deviation of distances to the closest intersection	11
	closest streets and intersections within 500m	e.g., total/average/standard deviation of distances to the closest intersection	11
Street-block	street-block itself	e.g., area, convexity, orientation, corner count, etc.	8
	buildings in blocks	e.g., count, total/average/standard deviation of the area of building	4
	street-block within 50, 200 and 500m	e.g., total/average/standard corner count, area of blocks etc.	14
Total features			129

3.3 Building Floor Estimation from Street-Level Images

Inspired by the work of automatic facade parsing in [20], we develop a building floor estimation workflow based on automatic facade parsing and urban architecture rules. In short, we aim to generate the estimation of building floor or height (by multiplying an average floor

height) as the “pseudo label” to guide ML regression models with the aforementioned OSM morphometric features as covariates. Figure 3 illustrates the developed method of building floor estimation based on SVI. To explain the developed method in more detail, we elaborate on three main steps as follows:



■ **Figure 3** Three steps of building floor estimation from street-level images: (1) aligning SVI and OSM building; (2) facade parsing using object detection; (3) generating “pseudo label” by building floor estimation.

SVI and OSM building alignment. As the first step, we download existing SVI from Mapillary via their open-source image API, where each SVI record consists of geotagged coordinates of the camera during a trip sequence and additional metadata information (Table 2), especially the compass angle of the camera direction (i.e., 0 to 360 degrees). This compass angle together with geotagged coordinates of the camera is key for aligning SVI with an individual OSM building. To this end, we apply a simple ray-tracing method to determine their relationship and assign the selected Mapillary SVI to its corresponding OSM building footprints (see Figure 3). Currently, we manually select Mapillary images which cover the complete facade of a building without being blocked by vegetation and cars, while future work is needed to automate this selecting process. A possible solution is to apply semantic segmentation approaches and ensure the skyline and ground are both visible within a single SVI.

Facade object detection. There are two common approaches in measuring building heights: either estimating absolute metrics (e.g., meters) or counting the floor number. As for accurately inferring the floor number, key features (e.g., window, balcony, and door) and their layout in the building facade play a key role [6]. Herein, we aim to detect these key features from street-level Mapillary imagery via the facade parsing technique. To this end, we follow the deep learning method developed in [20] for automatic facade parsing from the SVI data. Specifically, we use a pre-trained one-stage object detection network, namely YOLO

■ **Table 2** Selected metadata of SVI from the Mapillary Image API Endpoints.

Fields	Data Format	Description
computed_geometry	GeoJSON Point	latitude and longitude after running image processing.
computed_compass_angle	float	compass angle of the camera direction.
computed_altitude	float	altitude after running image processing, from sea level.
computed_rotation	enum	corrected orientation of the image, refer to OpenSfM definition.
camera_type	enum	type of camera projection: “perspective”, “fisheye”, “equirectangular”.
captured_at	timestamp	capture time of the camera.
camera_parameters	array of float	focal length, k1, k2 of the camera.
exif_orientation	enum	orientation of the camera as given by the Exif tag.

Note: All fields refer to Mapillary API Version 4.

v3 [31] (with the Darknet53 backbone), for the purpose of fast and accurate facade object detection. Herein, the facade object detection has been pre-trained on a facade semantic dataset called FaçadeWHU [20], thus could be directly applied to detect key facade features (e.g., window, balcony, and door) from the Mapillary SVI collected in Heidelberg without further training. As a result, the detected facade features are saved as a list of objects and their image coordinates.

Building floor estimation. Based on facade object detection results, we then apply a rule-based approach to determine the floor number in order to estimate the height of corresponding OSM buildings. Specifically, we first group facade objects (i.e., windows and doors) with their vertical coordinates and calculate the difference between each two neighbored elements, next k-mean clustering (with $k=2$) is used to find the clusters where objects are aligned vertically with each other, which results in a floor number estimation by counting the number of windows. By considering an average floor-to-floor height (i.e., 2.5 meters for residential or 3.5 meters for commercial), we can then derive the building height information from the SVI data, and use it as an SSL training label (h_{SSL}) to train the ML regression model on OSM morphometric features.

4 Preliminary Result

4.1 Case Study

As a case study, we implemented and tested the proposed method (Figure 2) in a classic western European city, namely the city of Heidelberg, Germany by considering Heidelberg was relatively well-mapped in OSM. Moreover, the reference data (h_{Raw}) of building heights obtained from the City of Heidelberg is also available, where building eaves heights (as we aim at LoD1 model for now) were recorded and spatially joined with OSM building footprints.

We extracted the latest OSM data (buildings and streets) via the ohsome API, which is built on the OpenStreetMap History Database (OSHDB) [30]. Herein, the ohsome API enables us to trace back to even historical OSM data, which can potentially contribute to more intrinsic features (e.g., the curve of nodes or contributions density). However, this goes beyond the scope of this paper. In this work, we calculated 129 morphometric features for 16,089 building footprints within the city of Heidelberg, which were used to train three types of ML regression models, specifically RF with 1000 trees, SVM with RBF kernel, and a three-layer dense CNN, to estimate building heights.

Regarding the SVI data, we followed the method described in Figure 3 by manually choosing 308 street-level Mapillary images and aligning them with 308 corresponding OSM building footprints by considering the SVI metadata. Then, we estimated their floor number

and further converted them into building heights by multiplying an average height of 2.5 meters for residential buildings and 3.5 meters for commercial and public buildings [9]. Herein, we manually verified the building function for these 308 SVI and their corresponding OSM building footprints. Although it is possible to automate this process with OSM data [13, 3], the prediction of building functions is beyond the scope of this paper. Despite its limitation, the proposed method provides a promising and low-cost solution to create open-source 3D city models (LoD1) by consuming only VGI data sourced (i.e., OSM and SVI) with a flexible SSL schema.

4.2 Experimental Result

In our case study, we conduct two comparative analysis to evaluate the capability of our SSL method w.r.t mainly two variables: first, the different OSM morphometric features, second, the different ratio of “pseudo label” during SSL training, by comparing the regress performance among three ML regression models (e.g., RF, SVM, and CNN).

Height estimation with different OSM features. To validate multi-level morphometric features extracted from OSM, Table 3 compares the regression performance of three ML models (RF, SVM, and CNN) using two different levels of morphometric features (i.e., 64 building-level features and all 129 features). Herein, we set a split ratio (between training and testing samples) of 0.7 on the reference data and calculate three common regression metrics (MAE, RMSE, and R^2 , all in meters) for the evaluation purpose. An important finding is that the integration of street and street-block features leads to an incremental boosting in the model performance, though this is less significant in the case of CNN. Though in the case of SVM, more features seem to be not helpful. A potential reason can be attributed to a potential effect of the curse of dimensionality. In short, an average MAE of around 2.3 meters (RF with 129 features), which is less than the average height of a single floor, confirms the feasibility of accurately estimating building height only from OSM morphometric features. This result encourages us to incorporate these OSM morphometric features with the proposed SSL method to better create large-scale and open-source 3D city models.

■ **Table 3** Preliminary results of estimating building heights with different OSM features and regression models.

	RF		SVM		CNN	
Feature	64	129	64	129	64	129
MAE	2.58	2.38	2.89	2.91	2.78	2.67
RMSE	3.55	3.34	3.89	3.91	3.71	3.62
R^2	0.2235	0.3140	0.0681	0.0567	0.1515	0.1929

SSL with different ratio of “pseudo label”. Based on the workflow described in Figure 3, we are able to collect 308 SVI from Mapillary and extract “pseudo label” via facade object detection, then associate these height values with their corresponding OSM building footprints. To test the impact of different SSL ratio, we set up three training sets: 1) to use only estimated heights from SVI (SVI) as an aggressive scenario of SSL; 2) to randomly select 308 OSM buildings and retrieve their heights from the reference data to simulate the fully supervised scenario (RAW); 3) to merge the “pseud label” with reference heights thus have a balance SSL training set (i.e., 308 each for SVI and RAW). In addition, a valuation

set with 2,000 buildings randomly extracted from the reference data is considered given the limited number of training labels. Table 4 shows the numerical results using different ratio of “pseudo label” (e.g., SVI, RAW, and SSL) and three ML regression models (with 129 features). Although the “pseudo label” (SVI) still leads to the largest error (w.r.t MAE and RMSE) in all three regression models, the “pseudo” height extracted from SVI is indeed informative for building height regression, more importantly, it is beneficial when merging with existing labels. Therefore, the quantitative result listed in Table 4 confirms that the proposed SSL method is effective and efficient in extracting “pseudo” training information from crowdsourced SVI data, which largely boosts the estimation accuracy using all three different ML regression models. In future work, it would be interesting to further investigate how different building types (e.g., residential or commercial, one-floor or multi-floor) can affect the accuracy of building height estimation.

■ **Table 4** Preliminary results of estimating building heights with different training sets and regression models.

Label	RF			SVM			CNN		
	SVI	RAW	SSL	SVI	RAW	SSL	SVI	RAW	SSL
MAE	2.75	2.67	2.07	2.93	2.89	2.20	3.23	3.03	2.72
RMSE	3.85	3.80	2.99	3.99	3.87	3.47	4.11	3.99	3.71
R^2	0.2302	0.2210	0.5368	0.1726	0.0315	0.3735	0.0241	0.1718	0.2458

Regarding the generation of “pseudo label”, Figure 4 shows selected examples of building floor estimations from Mapillary SVI in Heidelberg. One can observe that for lower floor numbers in case the captured facade is complete, the model works in a sensible way. However, we encountered several challenging cases when the building facade is not complete or the layout of windows (e.g., dormer windows) is difficult to be grouped by our floor estimation rules. In this context, future work is needed to develop a more robust method of extracting and distinguishing related features from SVI, such as roof types, dormer windows, and building functions, which can be helpful to generate more reliable “pseudo labels” for the SSL method.

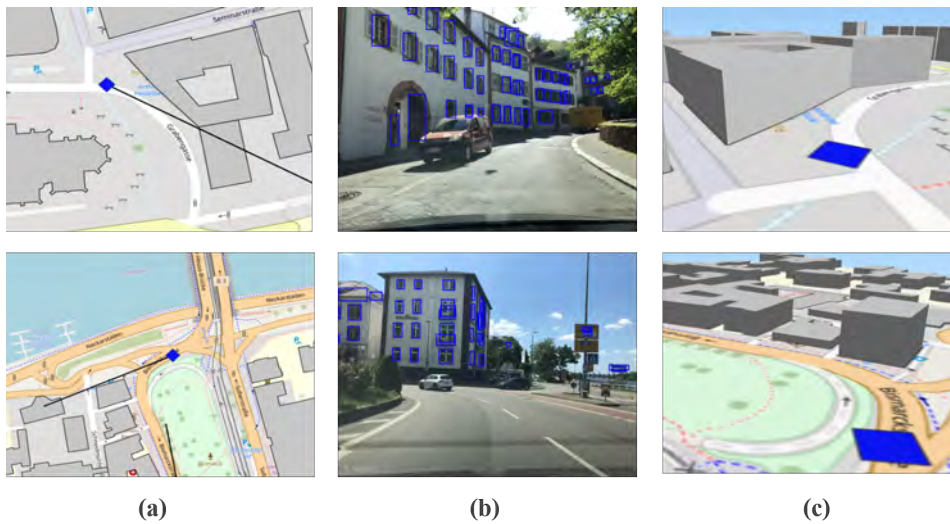
5 Discussion

In Figure 5, we demonstrate a 3D city model in LoD1 for selected buildings in the old town of Heidelberg, which is created using the proposed SSL method based on SVI (Figure 5 (b)) and OSM building footprints (Figure 5 (a)). In future work, we aim to refine this method by addressing the aforementioned limitations and comparing the estimated one with official LoD1 city models in selected cities.

Our preliminary result echoes the findings in [20] and [27] to a certain extent. More importantly, the SSL method will make our method in principle even more flexible and easy-to-apply in areas where the availability of training data (e.g., existing building heights) is limited or difficult-to-access. For instance, in most developed countries, 3D city models can be established using e.g., Digital Terrain Model (DTM), however the acquisition of large-scale and accurate DTM data remains costly and time-consuming. In this context, the proposed method provides a solution to directly harness existing crowdsourced VGI data (OSM and SVI) for 3D city modeling without additional data acquisition (e.g., DTM). Therefore, the “low-cost” herein mainly refers to the cost of traditional data acquisition methods w.r.t building height information. Despite the high potential, we identify several limitations to be addressed in future work:



■ **Figure 4** Selected examples of facade object detection and floor number estimating from Mapillary images in Heidelberg.



■ **Figure 5** The creation of a LoD1 3D city model using SVI and OSM data in the old town of Heidelberg. (a) OSM data with SVI metadata; (b) SVI with face object detection results; (c) LoD1 model with estimated building heights.

7:12 Automatic Building Height Estimation

- It is key to improve the building floor estimation workflow in terms of accuracy and speed, with which more “pseudo labels” can be extracted and used for SSL. For instance, the current SVI selection is done manually to ensure complete coverage of a building facade without being blocked by vegetation and cars, while this process can be automated using a semantic segmentation approach to improve the efficiency of generating high-quality “pseudo labels” at scale. Moreover, OSM data itself may contain information about building height (“*building:levels=* or height=**”) as well, which could be a helpful source to get more training data into the SSL method.
- Despite its low-cost and open-source nature, the quality aspect of VGI data (i.e., OSM and SVI data) remains under-quantified in this work, but certainly deserves a careful and decent treatment when applied to different countries or cities in the world [24]. For instance, the positional error and obstruction in SVI can significantly hinder the existing floor estimation approach. In addition, one needs to investigate how many SVI images are needed to have a reasonable spatial coverage of a study area to ensure the effectiveness of the SSL method.
- The ML regression models used in this work are based on a 1D vector feature space (up to 129 different features). However, a more sophisticated method is needed to encode the spatial relationship among buildings. For example, one option is to apply a graph CNN [38] as a spatial-explicit building height regressor.
- It is still unclear how different architecture types (e.g., roof type, construction age, building function) and city styles (e.g., low-rise, medium-rise, or high-rise) will affect the effectiveness and accuracy of our SSL method.

6 Conclusion

In this paper, we present a semi-supervised learning (SSL) method of automatic building height estimation by integrating crowdsourced street-level images (SVI) with multi-level morphometric features extracted from the OpenStreetMap (OSM) data. In this context, we design a workflow to convert facade object detection results from Mapillary SVI into “pseudo label” of building heights for three different ML regression models. As a case study, we validate the proposed SSL method in the city of Heidelberg, Germany, and the preliminary result looks very promising. However, the varying quality of volunteered geographical information (VGI) data, cultural and city-wise differences in the morphological features used, and the varying availability of SVI, all lead to certain limitations of such an SSL method. Our future work will focus on tackling these limitations and provide a robust and scalable solution of large-scale and open-source 3D city modeling purely based on low-cost VGI data.

References

- 1 A Alobeid, K Jacobsen, and C Heipke. Building height estimation in urban areas from very high resolution satellite stereo images. In *ISPRS Hannover Workshop*, volume 5, pages 2–5, 2009.
- 2 Joshua S Apte, Kyle P Messier, Shahzad Gani, Michael Brauer, Thomas W Kirchstetter, Melissa M Lunden, Julian D Marshall, Christopher J Portier, Roel CH Vermeulen, and Steven P Hamburg. High-resolution air pollution mapping with google street view cars: exploiting big data. *Environmental science & technology*, 51(12):6999–7008, 2017.
- 3 Kuldip Singh Atwal, Taylor Anderson, Dieter Pfoser, and Andreas Züfle. Predicting building types using openstreetmap. *Scientific Reports*, 12(1):19976, 2022.

- 4 Jérémy Bernard, Erwan Bocher, Elisabeth Le Saux Wiederhold, François Leconte, and Valéry Masson. Estimation of missing building height in OpenStreetMap data: A French case study using GeoClimate 0.0.1. *Geoscientific Model Development*, 15(19):7505–7532, October 2022. doi:10.5194/gmd-15-7505-2022.
- 5 Filip Biljecki and Yoong Shin Chow. Global building morphology indicators. *Computers, Environment and Urban Systems*, 95:101809, 2022.
- 6 Filip Biljecki, Hugo Ledoux, and Jantien Stoter. Generating 3d city models without elevation data. *Computers, Environment and Urban Systems*, 64:1–18, 2017.
- 7 Filip Biljecki, Hugo Ledoux, and JE Stoter. Height references of citygml lod1 buildings and their influence on applications. In *Proceedings. 9th ISPRS 3DGeoInfo Conference 2014, 11-13 November 2014, Dubai, UAE,(authors version)*. Citeseer, 2014.
- 8 Yinxia Cao and Xin Huang. A deep learning method for building height estimation using high-resolution multi-view imagery over urban areas: A case study of 42 Chinese cities. *Remote Sensing of Environment*, 264:112590, October 2021. doi:10.1016/j.rse.2021.112590.
- 9 Bumseok Chun and Jean-Michel Guldmann. Two- and three-dimensional urban core determinants of the urban heat island: A statistical approach. *Journal of Environmental Science and Engineering B*, 1(3):363–378, 2012.
- 10 Thomas Esch, Julian Zeidler, Daniela Palacios-Lopez, Mattia Marconcini, Achim Roth, Milena Mönks, Benjamin Leutner, Elisabeth Brzoska, Annekatrin Metz-Marconcini, Felix Bachofer, et al. Towards a large-scale 3d modeling of the built environment – joint analysis of tandem-x, sentinel-2 and open street map data. *Remote Sensing*, 12(15):2391, 2020.
- 11 Hongchao Fan and Liqiu Meng. A three-step approach of simplifying 3d buildings modeled by citygml. *International Journal of Geographical Information Science*, 26(6):1091–1107, 2012.
- 12 Hongchao Fan and Alexander Zipf. Modelling the world in 3d from vgi/crowdsourced data. *European handbook of crowdsourced geographic information*, 435, 2016.
- 13 Hongchao Fan, Alexander Zipf, and Qing Fu. Estimation of building types on openstreetmap based on urban morphology analysis. *Connecting a digital Europe through location and place*, pages 19–35, 2014.
- 14 Marcus Goetz. Towards generating highly detailed 3d citygml models from openstreetmap. *International Journal of Geographical Information Science*, 27(5):845–865, 2013.
- 15 Peng Gong, Zhan Li, Huabing Huang, Guoqing Sun, and Lei Wang. ICESat GLAS Data for Urban Environment Monitoring. *IEEE Transactions on Geoscience and Remote Sensing*, 49(3):1158–1172, March 2011. doi:10.1109/TGRS.2010.2070514.
- 16 Michael F. Goodchild. Citizens as sensors: The world of volunteered geography. *GeoJournal*, 69:211–221, August 2007. doi:10.1007/s10708-007-9111-y.
- 17 Krzysztof Janowicz, Song Gao, Grant McKenzie, Yingjie Hu, and Budhendra Bhaduri. Geoai: spatially explicit artificial intelligence techniques for geographic knowledge discovery and beyond. *International Journal of Geographical Information Science*, 34(4):625–636, 2020.
- 18 Longlong Jing and Yingli Tian. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 43(11):4037–4058, 2020.
- 19 Thomas H Kolbe, Gerhard Gröger, and Lutz Plümer. Citygml-3d city models and their potential for emergency response. In *Geospatial information technology for emergency response*, pages 273–290. CRC Press, 2008.
- 20 Gefei Kong and Hongchao Fan. Enhanced facade parsing for street-level images using convolutional neural networks. *IEEE Transactions on Geoscience and Remote Sensing*, 59(12):10519–10531, 2020.
- 21 Julia Kubanek, Eike-Marie Nolte, Hannes Taubenböck, Friedemann Wenzel, and Martin Kappas. Capacities of remote sensing for population estimation in urban areas. *Earthquake Hazard Impact and Urban Planning*, pages 45–66, 2014.

- 22 Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3(2), page 896, 2013.
- 23 Hao Li, Benjamin Herfort, Wei Huang, Mohammed Zia, and Alexander Zipf. Exploration of openstreetmap missing built-up areas using twitter hierarchical clustering and deep learning in mozambique. *ISPRS Journal of Photogrammetry and Remote Sensing*, 166:41–51, 2020.
- 24 Hao Li, Benjamin Herfort, Sven Lautenbach, Jiaoyan Chen, and Alexander Zipf. Improving openstreetmap missing building detection using few-shot transfer learning in sub-saharan africa. *Transactions in GIS*, 26(8):3125–3146, 2022.
- 25 Xuecao Li, Yuyu Zhou, Peng Gong, Karen C. Seto, and Nicholas Clinton. Developing a method to estimate building height from Sentinel-1 data. *Remote Sensing of Environment*, 240:111705, April 2020. doi:10.1016/j.rse.2020.111705.
- 26 Chao-Jung Liu, Vladimir A. Krylov, Paul Kane, Geraldine Kavanagh, and Rozenn Dahyot. IM2ELEVATION: Building Height Estimation from Single-View Aerial Imagery. *Remote Sensing*, 12(17):2719, January 2020. doi:10.3390/rs12172719.
- 27 Nikola Milojevic-Dupont, Nicolai Hans, Lynn H Kaack, Marius Zumwald, François Andrieux, Daniel de Barros Soares, Steffen Lohrey, Peter-Paul Pichler, and Felix Creutzig. Learning from urban form to predict building heights. *Plos one*, 15(12):e0242010, 2020.
- 28 Hui En Pang and Filip Biljecki. 3d building reconstruction from single street view images using deep learning. *International Journal of Applied Earth Observation and Geoinformation*, 112:102859, 2022.
- 29 Yujin Park and Jean-Michel Guldmann. Creating 3D city models with building footprints and LIDAR point cloud classification: A machine learning approach. *Computers, Environment and Urban Systems*, 75:76–89, May 2019. doi:10.1016/j.compenvurbsys.2019.01.004.
- 30 Martin Raifer, Rafael Troilo, Fabian Kowatsch, Michael Auer, Lukas Loos, Sabrina Marx, Katharina Przybill, Sascha Fendrich, Franz-Benjamin Mocnik, and Alexander Zipf. Oshdb: a framework for spatio-temporal analysis of openstreetmap history data. *Open Geospatial Data, Software and Standards*, 4:1–12, 2019.
- 31 Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.
- 32 Eirik Resch, Rolf André Bohne, Trond Kvamsdal, and Jardar Lohne. Impact of urban density and building height on energy use in cities. *Energy Procedia*, 96:800–814, 2016.
- 33 Yao Sun, Yuansheng Hua, Lichao Mou, and Xiao Xiang Zhu. Large-scale Building Height Estimation from Single VHR SAR image Using Fully Convolutional Network and GIS building footprints. In *2019 Joint Urban Remote Sensing Event (JURSE)*, pages 1–4, May 2019. doi:10.1109/JURSE.2019.8809037.
- 34 Heike Tost, Markus Reichert, Urs Braun, Iris Reinhard, Robin Peters, Sven Lautenbach, Andreas Hoell, Emanuel Schwarz, Ulrich Ebner-Priemer, Alexander Zipf, et al. Neural correlates of individual differences in affective benefit of real-life urban green space exposure. *Nature neuroscience*, 22(9):1389–1393, 2019.
- 35 Zhiyong Wang, Tessio Novack, Yingwei Yan, and Alexander Zipf. Quiet route planning for pedestrians in traffic noise polluted environments. *IEEE Transactions on Intelligent Transportation Systems*, 22(12):7573–7584, 2020.
- 36 Abraham Noah Wu and Filip Biljecki. Roofpedia: Automatic mapping of green and solar roofs for an open roofscape registry and evaluation of urban sustainability. *Landscape and Urban Planning*, 214:104167, 2021.
- 37 Michael Wurm, Hannes Taubenböck, Mathias Schardt, Thomas Esch, and Stefan Dech. Object-based image information fusion using multisensor earth observation data over urban areas. *International Journal of Image and Data Fusion*, 2(2):121–147, 2011.
- 38 Xiongfeng Yan, Tinghua Ai, Min Yang, and Hongmei Yin. A graph convolutional neural network for classification of building patterns using spatial vector data. *ISPRS journal of photogrammetry and remote sensing*, 150:259–273, 2019.

- 39 Yizhen Yan and Bo Huang. Estimation of building height using a single street view image via deep neural networks. *ISPRS Journal of Photogrammetry and Remote Sensing*, 192:83–98, October 2022. doi:10.1016/j.isprsjprs.2022.08.006.
- 40 Zhendong Yuan, Jules Kerckhoffs, Gerard Hoek, and Roel Vermeulen. A knowledge transfer approach to map long-term concentrations of hyperlocal air pollution from short-term mobile measurements. *Environmental Science & Technology*, September 2022. doi:10.1021/acs.est.2c05036.
- 41 Chaoquan Zhang, Hongchao Fan, and Gefei Kong. Vgi3d: an interactive and low-cost solution for 3d building modelling from street-level vgi images. *Journal of Geovisualization and Spatial Analysis*, 5(2):1–16, 2021.
- 42 Chenni Zhang, Yunfan Cui, Zeyao Zhu, San Jiang, and Wanshou Jiang. Building Height Extraction from GF-7 Satellite Images Based on Roof Contour Constrained Stereo Matching. *Remote Sensing*, 14(7):1566, January 2022. doi:10.3390/rs14071566.
- 43 Yunxiang Zhao, Jianzhong Qi, and Rui Zhang. CBHE: Corner-based Building Height Estimation for Complex Street Scene Images. In *The World Wide Web Conference, WWW '19*, pages 2436–2447, New York, NY, USA, May 2019. Association for Computing Machinery. doi:10.1145/3308558.3313394.
- 44 Xiaojin Jerry Zhu. *Semi-supervised learning literature survey*. University of Wisconsin-Madison Department of Computer Sciences, 2005.

Towards a Multidimensional Interaction Framework for Promoting Public Engagement in Citizen Science Projects

Maryam Lotfian¹ ✉ 

Institute INSIT, School of Business and Engineering Vaud, University of Applied Sciences and Arts Western Switzerland, Yverdon-les-Bains, Switzerland

Jens Ingensand ✉ 

Institute INSIT, School of Business and Engineering Vaud, University of Applied Sciences and Arts Western Switzerland, Yverdon-les-Bains, Switzerland

Christophe Claramunt ✉ 

Naval Academy Research Institute, Brest Naval, Lanveoc-Poulmic, BP 600, 29240 Brest Naval, France

Abstract

Citizen science (CS) projects are expanding into various fields and the number of CS applications is expanding. Despite this growth, engaging the public and sustaining their participation remains a challenge. Some studies have proposed that interacting with participants is an effective way to sustain their participation. This paper introduces a framework that outlines complementary levels of interaction including basic, incentivized, user-centered and action-oriented interactions. The interaction levels range from basic acknowledgments to instructions for taking action. The integration of these interactions within the spatial, temporal, and thematic dimensions is also discussed. The proposed framework is applied to a biodiversity CS project that involves different types of real-time feedback to participants based on the location, time, and image of the species observations. Location-based feedback is based on the species distribution models, and provides information on the probability of observing a certain species in a given location, as well as suggestions on the species to be observed in the participant's vicinity. Overall, the multi-dimensional interaction framework provides CS practitioners with insights into the various ways they can maintain communication with participants, whether through real-time machine-generated interactions or interactions between the project team and participants.

2012 ACM Subject Classification Human-centered computing → Collaborative content creation

Keywords and phrases Citizen Science, Multidimensional Interaction, Participation, User-centered Feedback, Machine Learning, Biodiversity

Digital Object Identifier 10.4230/LIPIcs.GIScience.2023.8

Supplementary Material *Software (Source Code)*: <https://github.com/mlotfian/Biosentiers-CS-functionality>; archived at [swh:1:dir:ea7f31c8ebd948814342017d44fe7d930b28db90](https://swh.1:dir:ea7f31c8ebd948814342017d44fe7d930b28db90)

Acknowledgements We would like to express our appreciation to the two reviewers for their invaluable and constructive feedback, which assisted us in improving the manuscript.

1 Introduction

Citizen science (CS), public participation in scientific projects [3], is not a new concept as for a long time amateur naturalists have been collecting animal and plant specimens and contributing to museum collections [24]. With rapid technological advancements in recent years, the number of CS projects has expanded significantly [19]. The advantage of technology-supported CS is that it favours advanced data collection processes and then additional

¹ corresponding author



interaction capabilities [20]. Despite the increased number of CS projects, the two essential features of engagement, initiating and sustaining participation, continue to be important concerns [31, 11]. Accordingly, several studies have been conducted to better understand the motivations of participants in contributing to CS projects [21, 8]. Consequently, the importance of interaction with participants and providing them feedback has been mentioned frequently as the main factors to keep citizens engaged [27, 18]. Nonetheless, few projects have investigated the role of feedback on increasing engagement, and if they have, the interaction has been primarily one-way and in the form of a generic response message such as an acknowledgment note or general information regarding the project [27]. Yet, less emphasis has been paid to user-centered feedback and interactions, or interactions that are specific to the contributions made by each participant. Biodiversity CS projects are an excellent example of user-centered feedback. Image recognition techniques are utilized to provide participants with the name of the species captured in their wildlife images [29]. This feedback helps them to verify the species name before uploading the observation to the platform, ensuring accuracy and promoting greater participation. In these projects, the feedback is centered around the participant's submitted image of the species. Although feedback based on images is valuable, even more accurate feedback can be achieved by considering additional factors such as the location and time of the observation. For instance, feedback could be provided on the probability of encountering a particular species at a specific location and time. Thus, considering various dimensions to interact with the participants is one very important element.

Another important factor to consider while interacting with participants is their heterogeneity. For example, for some participants, interaction means receiving incentives, for others, it means receiving acknowledgments and recognition, for yet others, it means active communication about the validity of their contributions and the project's progress, and finally for others, it means receiving guidance on taking actions that may even go beyond the project's objectives and/or time frame. As a result, it is essential to conceptualize the project in such a way that it accounts for these varying levels of interaction.

The objective of this research is to design and implement a participation platform that maximizes citizen interactions and encourages public engagement. Such interactions should not be limited to user single contribution but should include advanced capabilities that support different levels of platform feedback to the citizens, promoting active public engagement. These principles should be supported by practical user-friendly interfaces and applications experimented in real contexts. This paper introduces a framework that favours four levels of possible interactions with CS participants. Moreover, we investigate how these interactions are integrated within the three dimensions of space, time and theme. Furthermore, as a proof of concept, we present a case study of an implemented biodiversity CS project that shows how our approach may be put into practice, focusing on the third level of our interaction framework. In conclusion, we present the potential for expanding the case study and provide examples of the adaptability of the interaction framework to other CS applications in diverse fields.

2 Participation and communication in CS

Active public participation in CS projects can lead to the acquisition and contribution of knowledge, as well as the desire and satisfaction of being a part of a process and a community [33]. There are different categorizations of CS levels of participation, which are mainly focused on the degree of engagement in the project [18]. One of the most known classifications of participants is the one defined by Haklay [13]. Haklay's ladder

of participation includes four levels: crowdsourcing, distributed intelligence, participatory science, and extreme CS. As we progress through the levels, the usage of cognition in task performance increases, and participants become involved in greater phases of the project, such as in extreme CS where the participants are involved in problem definition, data collection, and analysis. In a later article [14], Haklay discussed the common conceptualizations of participation related to the misjudgment of participants based on their level of participation by categorizing low level participation (participating mainly to collect data) as “bad” and high level participation (participating in various phases of a project) as “good”. He designed a matrix with four cells where the participation to a project depends on the level of knowledge and level of engagement needed for the project, thus four possibilities of low/high, low/low, high/low, and high/high level of knowledge and level of engagement.

While the majority of the literature has been on the level of engagement and participation as measured by the extent to which people contribute to a CS project, less emphasis has been placed on the level of interactions with participants. Various CS projects focus primarily on obtaining data from participants rather than connecting with and understanding their needs as well as giving some information back to them, resulting in a failure to engage people to continue participating as well as a failure to learn from the project [9, 7, 17]. Accordingly, maintaining active communication with participants is critical, but what are the numerous methods by which scientists might develop this interaction between themselves and the citizens? Is it simply the sharing of information or showing appreciation? While communication is critical, some people do not need to engage with one another in order to contribute to a project, while others require active communication and information exchange [12]. It is thus important to understand how to define communication in CS.

Citizens can play different roles when being part of a participatory knowledge production process. Different levels of implications can be identified, from contributory to participatory and actor levels with citizens being progressively involved in dialogue-based relationships and empowerment [15]. At the abstract level, an interaction space generates a common framework for exchange using bilateral communications. Different dimensions can be identified to qualify such interaction space from the physical, spatio-temporal, semantics and technological dimensions. According to Hekker and Taddichen [15] communication in CS can take two forms. First, communication in its most fundamental sense, which is information exchange and two-way dialogues. Second communication as a tool, to identify and reach the target audience, to motivate participants to contribute, to negotiate interests, to provide feedback, and to communicate results.

Given that there are various types of CS projects [3], the goals of communication vary based on the type of project. For contributory projects, where scientists design the project and members of the public primarily contribute data, the goals of communication for citizens are to follow instructions, learn and apply them, and the goals of communication for scientists are to promote participation, increase motivation, and sustain participation [15]. However, in other types of projects where citizens are involved in more steps of the project, such as co-created projects, where the project is designed collaboratively by scientists and members of the public, the main goals of communication for citizens are to provide expertise, negotiate interests, exchange knowledge, create something together, and so on, and the goals of communication for scientists are similar to the citizens with the addition of managing conflicts among the partners [15].

For interaction with citizens, continuous attention is given to the different mechanisms that can foster participation and especially rewards offered to citizens [5, 6]. Certainly, citizens are more likely to get involved if they are convinced of the project’s importance.

8:4 Multidimensional Interaction Framework

Monetary rewards have been used and considered as a way to boost citizen participation [5], however, we believe that these methods may introduce biases in the participation process and the topic remains a subject of ongoing discussion in the scientific community. Online acknowledgement is indeed part of good practices but nevertheless they cannot generate further interactions. Rather than acknowledging or monetarizing them, citizens should be considered as active players that can significantly contribute to participatory projects, and giving them a sense of citizen contributors. Besides the studies on communication and interaction with citizens, categorization of these interactions taking into account multiple dimensions of space, time, and theme is missing, to the best of our knowledge.

3 Toward a multi-dimensional 4-level interaction framework

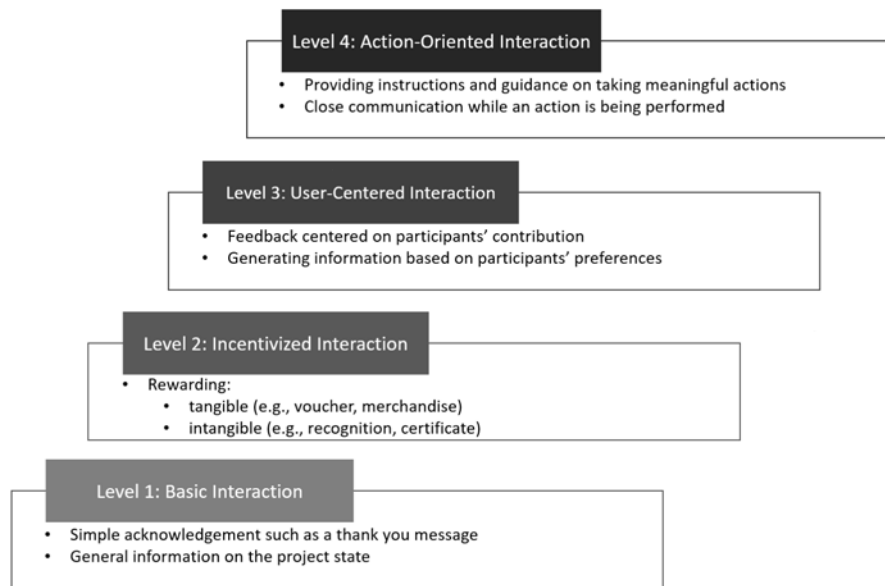
We introduce a multi-level framework for categorizing various levels of interaction with participants in CS projects, and then explain how these levels are integrated across the three dimensions of space, time, and theme. As there are diverse motivating factors to engage citizens in contributing to CS projects [21], some participants do not require any communication, while others demand active interaction to sustain their contribution. In this framework, illustrated in figure 1 four levels of interactions are defined: basic interaction, incentivized interaction, user-centered interaction, and action-oriented interaction.

Basic interaction: The first level includes basic feedback to the participants such as a thank you message after a contribution, or very general information about the state of the project. This level includes the participants whose input is independent of whether or not they are interacted with. They can be highly enthusiastic individuals who are passionate about a particular subject, such as bird watchers who are eager to report sightings and observations, as well as casual participants who may not be as actively engaged in the domain, but who are willing to contribute when the opportunity arises.

Incentivized interaction: This level includes interacting with the participants by rewarding them for their contribution. This interaction can take the form of rewards, which can be either tangible (such as vouchers) or intangible (such as points in a game or a certificate) [23]. These rewards are given in recognition of their contribution.

User-centered interaction: The third level includes providing feedback to participants, which are tailored to their contributions or in accordance to their preferences. This interaction level aims to maintain the contribution of participants that need to receive personalized feedback on their specific contributions. This feedback should not be general, but should rather be tailored to the individual's interests and preferences, taking into account the type of data they prefer to contribute, the location they prefer to contribute data from, and their preferred time for contributing data. In other words, the feedback should be user-centered and should address the three dimensions of space, time, and theme (combined or separated) to generate information that is specific to the participant.

Action-oriented interaction: This level of interaction involves providing participants with guidance and instructions that can be useful in performing an action. This level aims at maintaining the participants who are not only interested in receiving personalized feedback or useful information, but also desire guidance on ways they can actively contribute to the project's goals. For instance, in a biodiversity project, some participants may be more interested in finding out how they can help biodiversity, besides collecting observations. This can range from simple actions like planting a tree, to more elaborate measures like setting up bird feeders. These individuals seek guidance on actions they can take to assist with the project's objectives. The interaction at this level can go beyond the objective and time frame of the project depending on the type of actions that are proposed to the participants.



■ **Figure 1** The four level interaction framework. The interaction starts at a basic level, and as we progress through the levels, the focus of interaction shifts towards the contributions and requirements of the participants.

Furthermore, the four levels of interactions outlined before are incorporated within the framework of the three dimensions of space, time, and theme. Citizen observations are most frequently if not always built around the spatial (the where), temporal (the when) and thematic (the what) dimensions. In fact, spatial and temporal abstractions are fundamental to how humans perceive, conceptualize and experience their environment [26]. The respective roles of space and time have also been recognized in the development of interactive multimedia applications to both ensure contextual synchronization and then consistency [32]. For example, a given subject might have partial or complete knowledge of the spatial environment involved in multimedia interactions, as well as the one of the temporal coverage. A similar statement can be made regarding the thematic dimensions involved in such multimedia interactions. Indeed, in CS associated to geographical information, it makes sense to consider space, time and theme as fundamental information facets and structural dimensions to organise an interactive and active participatory framework.

When a participant makes a contribution, all the three dimensions converge since CS contributions occur in a specific location, at a certain time, and for a specified theme or subject. Figure 2 shows the connection between the three dimensions of space, time, and theme with various levels of interaction.

During the initial stage, all forms of participation are included within the convergence of the three dimensions to create a contribution. However, based on the requirements of the participant and the design of the project, the interaction with participants can be directed towards one specific dimension, two dimensions, or all three. Level 3 interaction falls under this category, where the interaction takes the form of user-centered feedback to the participant, and the feedback provided may vary based on the dimension.

The spatial dimension in a project can provide participants with location-based guidance, informing them of where a specific phenomenon is more likely to occur or what information can be collected at a particular location. However, some feedback is location-independent. For instance, the feedback can be tailored to the participant's preferred level of complexity

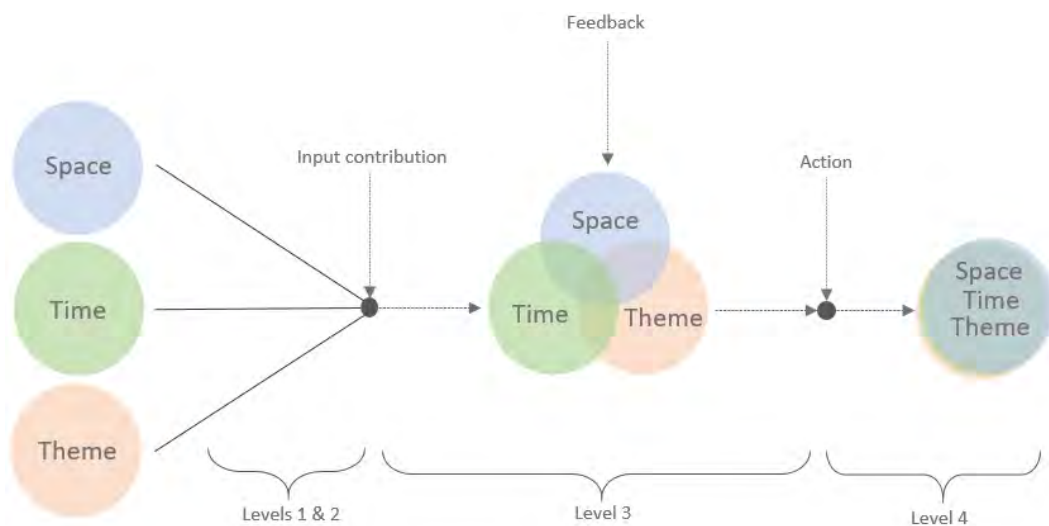
8:6 Multidimensional Interaction Framework

by proposing tasks related to their needs, thus customizing their contribution experience. In some cases, feedback can encompass multiple dimensions, including both spatial and temporal elements, giving the participant personalized information asking where and when they would like to make a contribution. These types of feedback are often associated with environmental data collection, where both the location and timing of the data collection play important roles. This type of feedback can allow participants to maximize the impact of their contributions and help ensure that the data collected is relevant and useful.

Finally, at the highest level of interaction where there is an action involved, all three dimensions converge again, as an action takes place at a specific location, time, and in relation to a specific theme, thus integrating all three dimensions.

Retroactions between the interaction framework and the citizens can be made at different levels of interactions, whether the priority is given to the spatial (e.g., species observed at a given location), temporal (e.g., species observed at a given time) or thematic dimension (e.g., where and when a given species is observed). This emphasizes the prominent role played by the spatial and temporal dimensions which are not limited to conventional attributes as they provide to the user specific capabilities for the selection and retroaction of data. In relation to level 4, and as done for the first and second levels, actions performed by a citizen are conducted at a given location and time, and for a particular theme and thus all dimensions converge.

The Next section presents a case study that focuses on the third level of the interaction framework and the three dimensions of space, time, and theme.



■ **Figure 2** The connection between three dimensions of space, time and theme before and after contribution to a citizen science project, and their relation to the levels of interaction (see Figure 1). When a contribution is made, all three dimensions merge and a general feedback can be provided (levels 1 and 2). Subsequently, based on the contribution, feedback may focus on one or a combination of dimensions (level 3). If participants receive feedback on how to perform an action (level 4) and carry it out, the three dimensions converge once more.

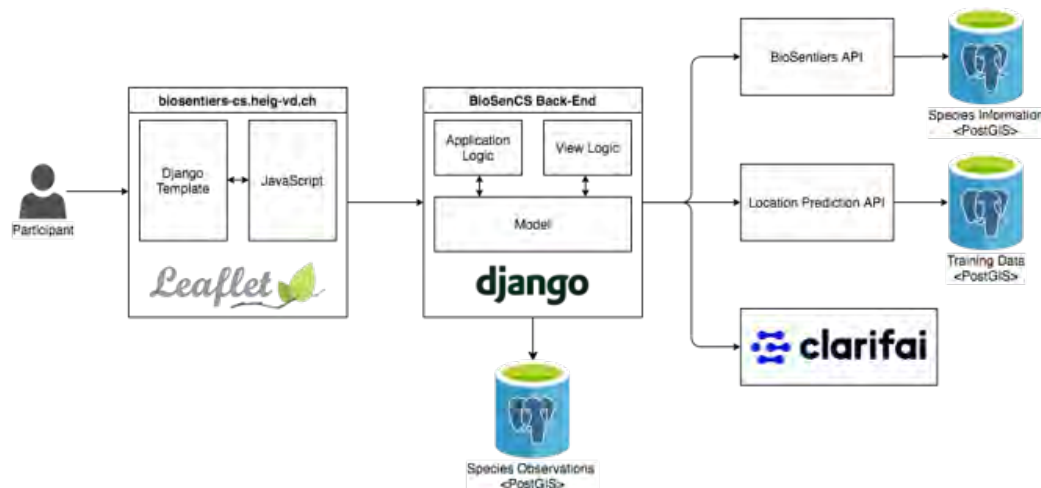
4 Case study

To delve deeper into the framework, we have carried out a case study that presents the interaction with participants within a biodiversity CS project. The case study presented here highlights three categories of feedback based on the dimensions of space, time, and theme. This feedback demonstrates the third level of interaction in our framework, illustrated in Figure 1, which entails providing feedback to participants in the form of personalized insights. The aim of this case study is to provide a concrete example of how the framework operates and how it can be applied in a real practice, serving as a proof of concept for its usefulness in other CS projects.

BioSenCS^{2,3} invites the public to collect biodiversity observations (with a focus on bird species) and at the same time applies automatic data validation to the observations while volunteers contribute and provides them with real-time user-centered feedback[22]. The main objectives of BioSenCS in relation to the third level of the interaction framework are as follows:

- Provide participants with real-time feedback based on the location, time, and image of the species observation
- Boost public engagement as a result of user-centered feedback
- Provide a learning opportunity for the participants through the feedback
- Enhance data quality through learning from automatic feedback

BioSenCS is implemented using the Django framework⁴, which is a Python-based free and open-source web framework, and we used a PostgreSQL⁵/PostGIS⁶ database for constructing our data models and preserving the collected observations. The high-level architecture of BioSenCS application is illustrated in figure 3.



■ **Figure 3** The high-level architecture of BioSenCS application.

² <https://biosentiers-cs.heig-vd.ch/>

³ <https://github.com/mlotfian/Biosentiers-CS-functionality>

⁴ <https://www.djangoproject.com/>

⁵ <https://www.postgresql.org/>

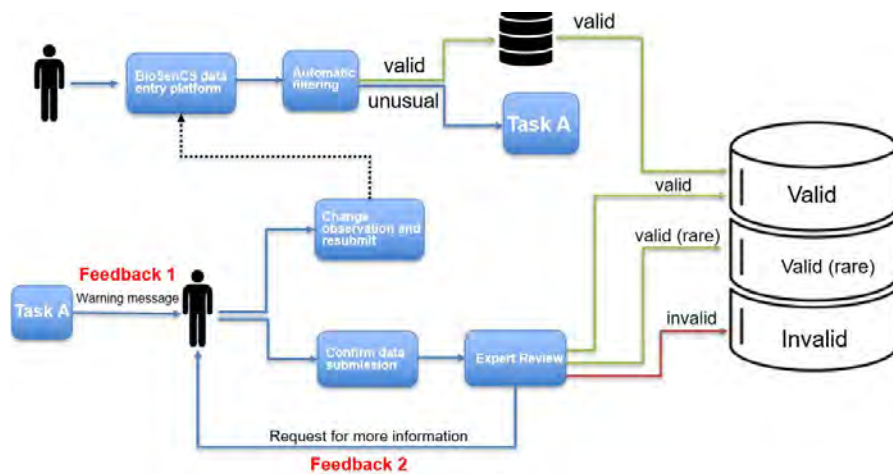
⁶ <https://postgis.net/>

The primary goals of this case study are to utilize an automated system to validate or filter observations and to provide real-time feedback to participants. The process involves an automatic filtering system, as depicted in Figure 4. When a participant submits an observation, it is first evaluated by the automatic system. If the observation does not meet certain criteria, it is marked as an unusual observation and the participant is provided with feedback including a detailed explanation of why the observation was flagged, which the feedback is based on any dimension of space, time, and theme separately or combined.

At this point, the participants have two options. They can either make changes to the observation based on the feedback received, or they can choose to proceed with the original submission, moving the observation to the final expert validation stage. If the expert determines that more information is needed, they will provide additional feedback to the participant.

The automatic validation process includes three elements: date validation, image validation, and location validation. The image and location validation utilize machine learning (ML) algorithms, while the date validation is performed by comparing the observation dates to a static dataset provided by ecologists. The dates dataset is accessible through the BioSentiers API (Application Programming Interface) (Figure 3).

Location validation is only applied to bird species, but image and date validation are applied to all four organisms: bird, butterfly, tree, and flower. The following sections concentrate on the third level of the interaction framework (See Figure 1) and integrate the three dimensions (See Figure 2) of theme, time, and space respectively, in the order they are presented.



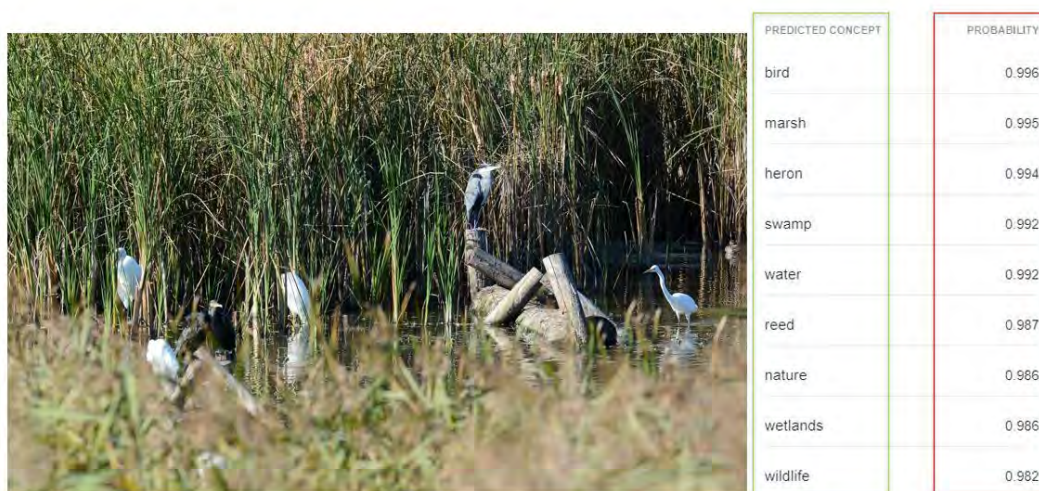
■ **Figure 4** The automatic data validation procedure applied in BioSenCS.

4.1 Theme: Observation Image Verification

The image filtering screens the contributed images that do not include the reported species (bird, flower, tree, or butterfly). The feedback obtained in image filtering focuses on the theme dimension of the observation, specifically the type of species. To perform image filtering, we used an artificial intelligence (AI) platform called Clarifai⁷. Clarifai is an AI company

⁷ <https://www.clarifai.com/>

that specializes in computer vision. It provides pre-trained models⁸ as well as the option of training a model with a custom dataset. Clarifai offers services via its API (1000 free API calls per month), which has a fast response time and can be integrated into AI-powered mobile or web applications. We used its general model⁹ to determine, for example, whether an image with bird tag really contains a bird or not. Once an image is sent to Clarifai's API, the model generates a set of possible tags that are present in the image along with their probability scores (See figure 5). We flagged an observation and sent a feedback to the participant, if the probability of having the species in the uploaded image was less than 85 percent. Figure 6 (b) illustrates the real-time image feedback to the participants.



■ **Figure 5** An example of Clarifai predicted tags and their probabilities for an observation contributed to BioSenCS.

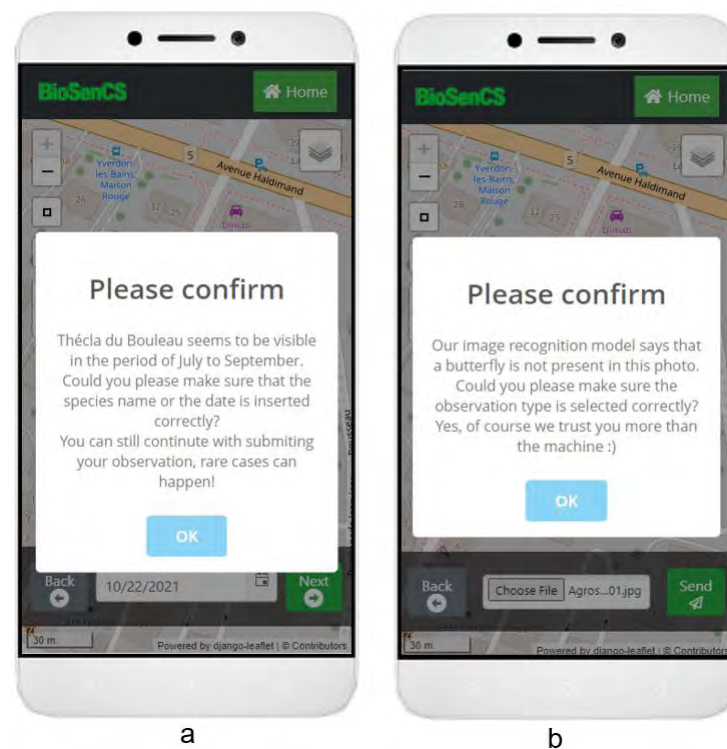
4.2 Time: Observation Date Verification

The date filter focuses on the time dimension of the observation and verifies whether or not, given a species name, the observation date falls within the species visibility period (the time where the species is mostly probable to be observed). Accordingly, we have used a dataset, which includes information of the visibility period of the species and is accessible through an API called BioSentiers¹⁰[16], and the observation date is verified using the two attributes of `periodStart` and `periodEnd` in the database. If the observation date is outside the species visibility period, the observation is considered as an outlier and the participant receives a feedback with information about the months (or the periods) the species can normally be observed, and asking the participant to verify the added observation (e.g. species name or the date). The final decision is however given to the participant, and the participant is not forced to modify the observation. The observation is however flagged in our database in a boolean attribute `flagDate` to be verified by experts later on. Figure 6 (a) illustrates the real-time date feedback to the participant.

⁸ <https://www.clarifai.com/developers/pre-trained-models>

⁹ <https://www.clarifai.com/models/image-recognition-ai>

¹⁰ <https://biosentiers.heig-vd.ch/api/species>



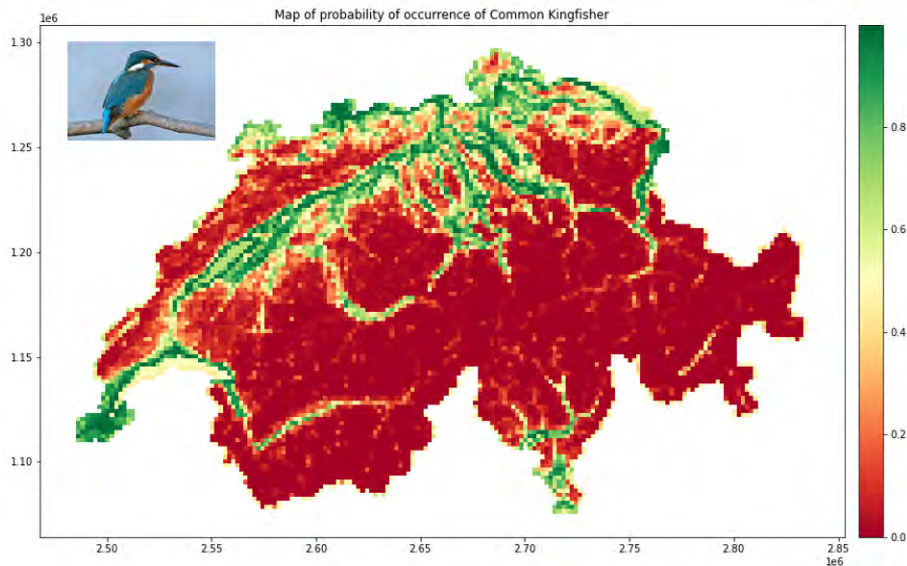
■ **Figure 6** Automatic date (a) and image (b) feedback in BioSenCS application.

4.3 Space: Observation Location Verification

Location verification focuses on the space dimension of the observation, and thus the corresponding feedback is centered on the participant's location. To perform location validation, we determined how the environmental variables surrounding the observation location corresponded to the species habitat characteristics. To accomplish this, we generated models of the distribution of the species in relation to the environmental variables in our study area (Switzerland). Accordingly, we used species distribution modeling (SDM) techniques [10] for bird species in Switzerland. SDM is a class of numerical models that explain how the presence or absence of a species at a given location is related to environmental (e.g. temperature, precipitation, etc.) and landscape characteristics (e.g. land cover, elevation, slope, etc.) [10]. These techniques are used to gain ecological and evolutionary insights as well as to predict distributions across landscapes, which requires spatial and/or temporal extrapolation. SDM can be used to understand how a species' distribution is correlated with its location, as well as to predict the locations of species occurrence where no data is available. To generate SDM two important datasets are required: the species abundance data and the environmental variables.

For the species dataset, we used eBird data [28] for Switzerland from 2016 to 2020, and for the environmental variables we used land cover (to obtain landscape proportions such as the percentage of forest, water bodies, etc. in a given area), elevation, slope, and NDVI (Normalized Difference Vegetation Index). Additionally, to generate SDMs, four algorithms were trained and compared based on their performance: Naive Bayesian (NB)[30], Random Forest (RF)[4], Balanced Random Forest (B-RF)[2], and a Deep Neural Network (DNN)[1]. The models trained with Balanced-RF performed better compared to the other three algorithms, and thus they were used to verify the location of new contributed observations.

For each species, we obtained two output distributions maps: a binary classification, and a map of probability of occurrence of the species over the whole of Switzerland. Figure 7 illustrates the map of probability of occurrence for Common kingfisher species. As shown in the figure, Common kingfisher can be mainly observed in the northern, and north west parts of Switzerland with high probability to be observed near lakes and water bodies.



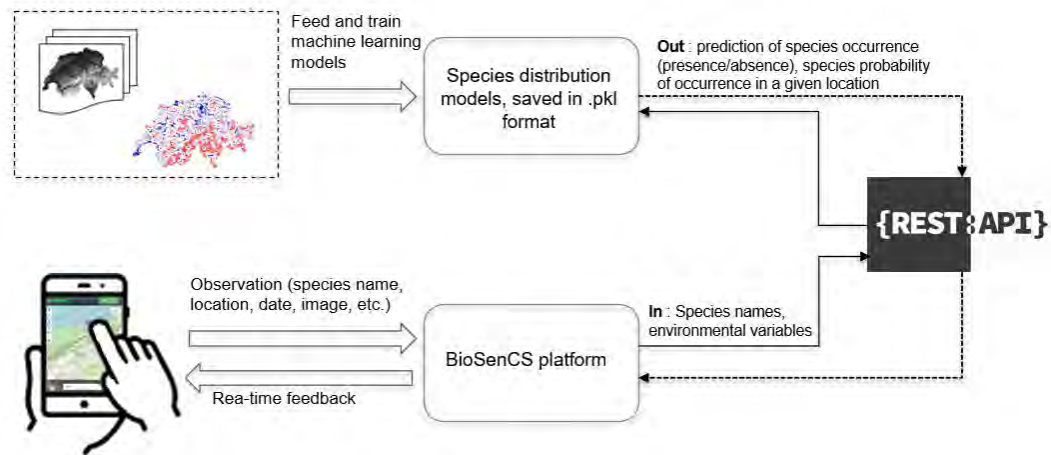
■ **Figure 7** Classification of probability of occurrence of Common kingfisher.

Once training the algorithms and choosing the best performant one, in this case Balanced-RF, the trained models were saved to be used for validating new contributed observations. We developed an API called BioLocation, that uses the trained models and that is integrated to the BioSenCS application to validate new observations while also providing user-centered suggestions on the top-five high-probable species that can be observed around the participant's location. The API takes the species name and location and returns the probability of observing the species in a $2km^2$ neighbourhood around the given location. The probability is given back to the participant as a real time feedback with information on species habitat characteristics. Additionally, the API can take the location and suggests the possible species that can be observed in the participant's proximity. Figure 8 illustrates the process of real time feedback generation taking into account the space dimension.

If the probability of observing a species in a particular location is higher than 50 percent (to account for randomness, as agreed when implementing the project), the generated feedback will simply provide complementary information to the participant, such as the possible places where the species is more probable to be observed. However, if the probability is less than 50 percent, the participant will be asked to confirm the validity of the observation (either the location or the species name). After receiving the feedback, the participant has the option to alter the observation based on the given information or leave it unchanged. Figure 9, a and b illustrate the two possible location feedback, and c illustrates the top five species probable to be observed around the location of the participant.

The feedback generated in this case study either focuses solely on one dimension, such as only the observation date or location, or a combination of dimensions, such as in the user-centered species suggestion which integrates space (user location) and theme (species names). Furthermore, when a participant encounters an unknown species, the top five

8:12 Multidimensional Interaction Framework

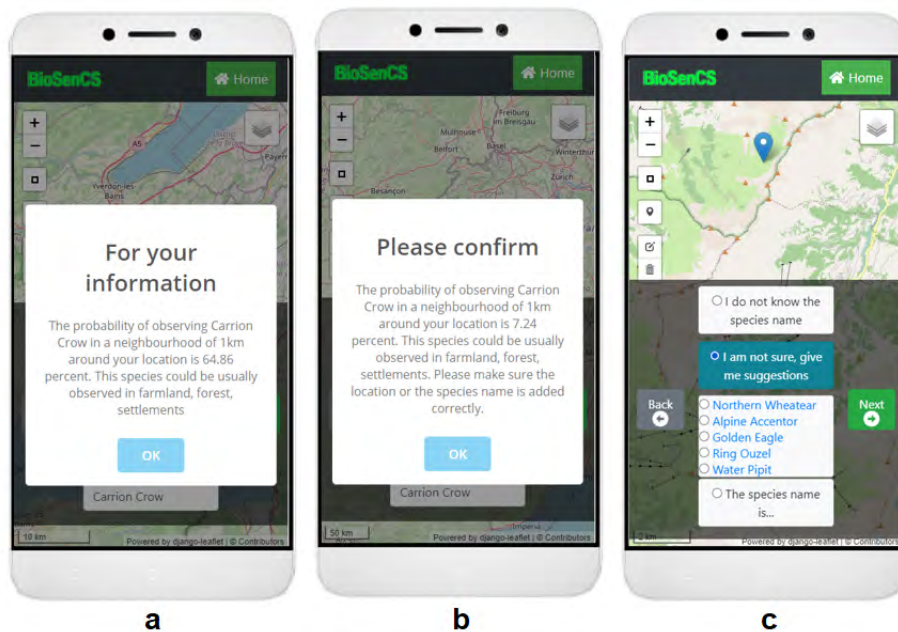


■ **Figure 8** Real time location feedback generation.

suggestions can assist in identifying the species from the list, thus again a combination of space and theme dimensions. Additionally, the feedback on the time dimension can provide information on the visibility period of different species, allowing participants to determine when they can observe species of interest, thus combining the time and theme dimensions.

Finally, a user test was conducted to gather feedback on the BioSenCS application interface and explore participant views on receiving automatic feedback. The application was promoted through targeted emails sent directly to students and colleagues within our university, as well as through social media platforms such as Facebook and LinkedIn. Additionally, word of mouth played a role in spreading awareness about the application. A thorough testing phase lasting three weeks was conducted, resulting in 224 visits to the application. Out of these visitors, 38 users successfully created an account, and 14 individuals actively participated in collecting observations. Additionally, during this three-week timeframe, a total of 230 observations were collected.

Following the completion of the testing period, participants were provided with a questionnaire that encompassed general inquiries about the application's usability. Additionally, participants were asked to evaluate the extent to which they found the feedback information useful and whether receiving feedback heightened their motivation to contribute to the project. These questions were rated on a 5-point Likert scale, with 1 representing "not at all useful/motivating" and 5 denoting "very useful/motivating." The average score for the usefulness of feedback was 3.33, while the average score for the impact of feedback on motivation was 3.5. To assess the impact of feedback on enhancing data quality, we examined the relationship between the number of flagged observations (O_F) and the total number of contributed observations (O_T) per user. By analyzing the correlation between the ratio of flagged observations to the total number of observations (O_F/O_T) and O_T , we discovered a statistically significant negative correlation of -0.63 (p-value = 0.036). This indicates that participants who made a higher number of contributions had fewer flagged observations. Essentially, this suggests that participants either utilized the feedback provided to improve their observations before submitting them (e.g., verifying accurate location pin placement) or developed the ability to provide higher quality data. Although we did not obtain explicit statistical evidence regarding the correlation between contribution over time and data quality, the findings imply that increased participant contributions lead to a reduced number of flagged observations, thus indicating higher quality data. However, a longer testing period would enable a clearer understanding of how feedback influenced data quality over time.



■ **Figure 9** Location feedback if probability of occurrence of species is higher (a) and lower (b) than 50%, and user-centered suggestion (c).

5 Discussion and conclusion

CS projects are rapidly expanding into various fields and the number of applications is growing, thanks to technological advancements [25]. Despite this growth, there is still a challenge in engaging the public to participate. One approach to sustaining participants' engagement is through interaction with the participants [15]. Accordingly, some projects provide general feedback to the participants, such as an acknowledgment of their contribution, or feedback focused on one aspect of their participation. However, the needs and preferences of participants may vary and require personalized feedback. This paper introduces a framework that categorizes different types of interactions with participants while considering three dimensions of space, time, and theme, when interacting with them for a more effective outcome.

The framework outlines four levels of interaction, beginning with the simplest form, such as acknowledging the participant, then is incentivized interaction offering tangible benefits like certificates or rewards to participants. The next level, user-centered interaction, is centered around the participant and involves two-way interaction, where feedback is tailored to their contributions and they are given the opportunity to interact with the project and express their preferences. The final level is action-oriented, providing instructions, guidance, and support to help the participant take meaningful actions. Additionally, the integration of the interaction levels within the three dimensions of space, time, and theme is discussed in this article. The feedback in level 3 can concentrate on one dimension or a combination of all three, while in level 4 interaction, all dimensions are combined, as an action is carried out at a particular location and time, and for a specific theme.

The biodiversity CS case study highlighted in this article emphasizes user-centered interaction and focuses on the third level of the interaction framework. The feedback provided is based on location, time, and images of the observations. The participants receive

8:14 Multidimensional Interaction Framework

location-based feedback on the likelihood of observing a specific species at a given location, and also receive suggestions for species to observe in their vicinity, combining both spatial and thematic dimensions. Although this case study provides multi-dimensional user-centered feedback, to fully realize its potential, a two-way interaction between participants and the project is necessary. To accomplish this, the three main questions of “where,” “when,” and “what” can help guide participants to receive targeted feedback. The following are examples of how the feedback in the case study can be expanded in each of the three dimensions.

- Spatial
 - Ask participants where they prefer to observe species.
 - Find out what types of environments participants prefer for collecting observations.
 - Based on their previous observations, suggest other locations they might be interested in visiting.
 - If participants have a particular species in mind, suggest likely locations where they can observe it.
- Temporal
 - Ask participants when they like to collect observations.
 - If they have a specific species in mind, recommend the best times to observe it.
 - Identify species that can be observed during the given time frame.
- Thematic
 - Ask participants what types of species they like to observe.
 - Find out what individual species they prefer to observe.
 - Based on their history of observations, suggest other species they might also be interested in observing.

The guidance questions mentioned above for generating user-centered feedback are focused on the biodiversity field, however, they can be adapted to other areas depending on the project’s objectives and the participants’ preferences. For instance, in a CS urban planning project aimed at mapping street and sidewalk quality, the spatial dimension could involve identifying and proposing areas with data gaps near participants, inviting them to collect data in such areas, and providing them with information on the quality of streets and sidewalks along their planned route.

Additionally, BioSenCS aims to expand by offering interaction at the fourth level (action-oriented) of the proposed framework in this article. This interaction can include providing support for preserving biodiversity such as through recommendations on appropriate plant species to grow in specific locations and at specific time frames, or offering guidance on building a garden pond, including information on necessary materials, estimated time required, and the best time to start construction based on the participant’s location. The fourth level interaction involves not only providing information but also maintaining a connection with the participants throughout the entire action-taking process. This connection can be through either automated and machine interactions, such as verifying the actions to be taken given a location and time, or through online or in-person interaction with the project team.

Overall, this article aims to present ways in which CS practitioners can interact with their participants, taking into account spatial, temporal, and thematic dimensions. The goal of categorization of interaction levels presented in this article is not to assign positive or negative labels to the levels, but rather to provide different methods to maintain community involvement in CS projects. The choice of interaction methods can be adjusted based on the project’s goals, its timeline, the preferences of the participants, and may consist of a single approach or a combination of approaches. Finally, the purpose of this article is not to give detailed instructions on how to interact with or provide feedback to the participants, but rather to present an overall perspective, supported by a relevant case study.

References


- 1 Oludare Isaac Abiodun, Aman Jantan, Abiodun Esther Omolara, Kemi Victoria Dada, Nachaat AbdElatif Mohamed, and Humaira Arshad. State-of-the-art in artificial neural network applications: A survey. *Heliyon*, 4(11):e00938, 2018. doi:10.1016/j.heliyon.2018.e00938.
- 2 Zahra Putri Agusta et al. Modified balanced random forest for improving imbalanced data prediction. *International Journal of Advances in Intelligent Informatics*, 5(1):58–65, 2019.
- 3 Rick Bonney, Heidi Ballard, Rebecca Jordan, Ellen McCallie, Tina Phillips, Jennifer Shirk, and Candie C Wilderman. Public participation in scientific research: Defining the field and assessing its potential for informal science education. a caise inquiry group report. *Washington D.C.: Center for Advancement of Informal Science Education (CAISE)*, 2009.
- 4 Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.
- 5 Francesco Cappa, Jeffrey Laut, Maurizio Porfiri, and Luca Giustiniano. Bring them aboard: Rewarding participation in technology-mediated citizen science projects. *Computers in Human Behavior*, 89:246–257, 2018.
- 6 Francesco Cappa, Federica Rosso, and Darren Hayes. Monetary and social rewards for crowdsourcing. *Sustainability*, 11(10), 2019. doi:10.3390/su11102834.
- 7 Sarah Composto, Jens Ingensand, Marion Nappez, Olivier Ertz, Daniel Rappo, Rémi Bovard, Ivo Widmer, and Stéphane Joost. How to recruit and motivate users to utilize vgi-systems? In *Proceedings of 19th AGILE International Conference on Geographic Information Science, 14-17th June 2016, Helsinki, Finland*, 2016.
- 8 Vickie Curtis. *Online citizen science projects: an exploration of motivation, contribution and participation*. Open University (United Kingdom), 2015.
- 9 Caroline Gottschalk Druschke and Carrie E. Seltzer. Failures of engagement: Lessons learned from a citizen science pilot study. *Applied Environmental Education & Communication*, 11(3-4):178–188, 2012. doi:10.1080/1533015X.2012.777224.
- 10 Jane Elith and John R. Leathwick. Species distribution models: Ecological explanation and prediction across space and time. *Annual Review of Ecology, Evolution, and Systematics*, 40:677–697, December 2009. doi:10.1146/annurev.ecolsys.110308.120159.
- 11 Glyn Everett and Hilary Geoghegan. Initiating and continuing participation in citizen science for natural history. *BMC ecology*, 16(1):15–22, 2016.
- 12 Hilary Geoghegan, Alison Dyke, Rachel Pateman, Sarah West, and Glyn Everett. Understanding motivations for citizen science. *Final report on behalf of UKEOF, University of Reading, Stockholm Environment Institute (University of York) and University of the West of England*, 2016.
- 13 Muki Haklay. Citizen science and volunteered geographic information: Overview and typology of participation. *Crowdsourcing geographic knowledge*, pages 105–122, 2013.
- 14 Muki Haklay et al. Participatory citizen science. *Citizen science: Innovation in open science, society and policy*, pages 52–62, 2018.
- 15 Susanne Hecker and Monika Taddicken. Deconstructing citizen science: a framework on communication and interaction using the concept of roles. *Journal of Science Communication*, 21(1):A07, 2022.
- 16 Jens Ingensand, Maryam Lotfian, Olivier Ertz, David Piot, Sarah Composto, Mathias Oberson, Simon Oulevay, and Mélanie Da Cunha. Augmented reality technologies for biodiversity education. In *Proceedings of the 21st Conference on Geo-information science, AGILE, Lund, Sweden*. 12-15 June, 2018.
- 17 Jens Ingensand, Marion Nappez, Stéphane Joost, Ivo Widmer, Olivier Ertz, and Daniel Rappo. The urbangene project: Experience from a crowdsourced mapping campaign. In *2015 1st International Conference on Geographical Information Systems Theory, Applications and Management (GISTAM)*, pages 1–7, 2015.
- 18 Anne Land-Zandstra, Gaia Agnello, and Yaşar Selman Gültekin. Participants in citizen science. *The science of citizen science*, 243, 2021.

- 19 Anne M. Land-Zandstra, Jeroen L. A. Devilee, Frans Snik, Franka Buurmeijer, and Jos M. van den Broek. Citizen science on a smartphone: Participants' motivations and learning. *Public Understanding of Science*, 25(1):45–60, 2016. doi:10.1177/0963662515602406.
- 20 Rob Lemmens, Vyrion Antoniou, Philipp Hummer, and Chryssy Potsiou. *Citizen Science in the Digital World of Apps*, pages 461–474. Springer International Publishing, Cham, 2021. doi:10.1007/978-3-030-58278-4_23.
- 21 Maryam Lotfian, Jens Ingensand, and Maria Antonia Brovelli. A framework for classifying participant motivation that considers the typology of citizen science projects. *ISPRS International Journal of Geo-Information*, 9(12):704, 2020.
- 22 Maryam Lotfian, Jens Ingensand, Olivier Ertz, Simon Oulevay, and Thibaud Chassin. Auto-filtering validation in citizen science biodiversity monitoring. In *Proceedings of the ICA; Proceedings of 29th International Cartographic Conference, Tokyo, Japan. 15-20 July, 2019*.
- 23 Michael Meder, Till Plumbaum, Aleksander Raczkowski, Brijnesh Jain, and Sahin Albayrak. Gamification in e-commerce: Tangible vs. intangible rewards. In *Proceedings of the 22nd International Academic Mindtrek Conference, Mindtrek '18*, pages 11–19, New York, NY, USA, 2018. Association for Computing Machinery. doi:10.1145/3275116.3275126.
- 24 Abraham Miller-Rushing, Richard Primack, and Rick Bonney. The history of public participation in ecological research. *Frontiers in Ecology and the Environment*, 10(6):285–290, August 2012. doi:10.1890/110278.
- 25 Greg Newman, Andrea Wiggins, Alycia Crall, Eric Graham, Sarah Newman, and Kevin Crowston. The future of citizen science: emerging technologies and shifting paradigms. *Frontiers in Ecology and the Environment*, 10(6):298–304, 2012.
- 26 Rafael Núñez and Kensy Cooperrider. The tangle of space and time in human cognition. *Trends in Cognitive Sciences*, 17(5):220–229, May 2013. doi:10.1016/j.tics.2013.03.008.
- 27 Simone Rüfenacht, Tim Woods, Gaia Agnello, Margaret Gold, Philipp Hummer, Anne Land-Zandstra, and Andrea Sieber. Communication and dissemination in citizen science. *The Science of Citizen Science*, 475:520, 2021.
- 28 Brian L Sullivan, Christopher L Wood, Marshall J Iliff, Rick E Bonney, Daniel Fink, and Steve Kelling. ebird: A citizen-based bird observation network in the biological sciences. *Biological conservation*, 142(10):2282–2292, 2009.
- 29 René van der Wal, Nirwan Sharma, Chris Mellish, Annie Robinson, and Advait Siddharthan. The role of automated feedback in training and retaining biological recorders for citizen science. *Conservation Biology*, 30(3):550–561, April 2016. doi:10.1111/cobi.12705.
- 30 Geoffrey I Webb, Eamonn Keogh, and Risto Miikkulainen. Naïve bayes. *Encyclopedia of machine learning*, 15:713–714, 2010.
- 31 Sarah Elizabeth West and Rachel Mary Pateman. Recruiting and retaining participants in citizen science: what can be learned from the volunteering literature? *Citizen Science: Theory and Practice*, 2016.
- 32 Wanmin Wu, Ahsan Arefin, Raoul Rivas, Klara Nahrstedt, Renata Sheppard, and Zhenyu Yang. Quality of experience in distributed interactive multimedia environments: toward a theoretical framework. In *Proceedings of the 17th ACM international conference on Multimedia*, pages 481–490, 2009.
- 33 Walter W Wymer Jr. Differentiating literacy volunteers: A segmentation analysis for target marketing. *International Journal of Nonprofit and Voluntary Sector Marketing*, 8(3):267–285, 2003.

Platial k -Anonymity: Improving Location Anonymity Through Temporal Popularity Signatures

Grant McKenzie ✉ 🏠 

Platial Analysis Lab, McGill University, Montréal, Canada

Hongyu Zhang ✉ 🏠 

Platial Analysis Lab, McGill University, Montréal, Canada

Abstract

While it is increasingly necessary in today's digital society, sharing personal location information comes at a cost. Sharing one's precise place of interest, e.g., Compass Coffee, enables a range of location-based services, but substantially reduces the individual's privacy. Methods have been developed to obfuscate and anonymize location data while still maintaining a degree of utility. One such approach, spatial k -anonymity, aims to ensure an individual's level of anonymity by reporting their location as a set of k potential locations rather than their actual location alone. Larger values of k increase spatial anonymity while decreasing the utility of the location information. Typical examples of spatial k -anonymized datasets present elements as simple geographic points with no attributes or contextual information. In this work, we demonstrate that the addition of publicly available contextual data can significantly reduce the anonymity of a k -anonymized dataset. Through the analysis of place type temporal visitation patterns, hours of operation, and popularity values, one's anonymity can be decreased by more than 50 percent. We propose a platial k -anonymity approach that leverages a combination of temporal popularity signatures and reports the amount that k must increase in order to maintain a certain level of anonymity. Finally, a method for reporting platial k -anonymous regions is presented and the implications of our methods are discussed.

2012 ACM Subject Classification Security and privacy → Privacy protections; Information systems → Location based services; Information systems → Geographic information systems

Keywords and phrases location anonymity, location privacy, geoprivacy, place, temporal, geosocial

Digital Object Identifier 10.4230/LIPICs.GIScience.2023.9

Supplementary Material *Other (Code)*: <https://github.com/ptal-io/platial-k-anonymity>
archived at `swh:1:dir:d359af2244e4dc123d656fe613a9c2a3d3d6f985`

Funding Fonds de Recherche du Québec – Société et culture (Award Number NP-281897).

1 Introduction

In 2014, a student used time-stamped paparazzi photographs of celebrities exiting taxicabs in New York City (NYC) to identify their home locations using a supposedly anonymized dataset of taxicab trips [34]. This raised privacy concerns about the dataset [9] and forced the NYC Taxi & Limousine Commission to revisit their anonymization process and obfuscate trip origins and destinations in latter data releases. The lesson to be learned from this privacy debacle is that even though a dataset may have been anonymized, it does not exist in a vacuum. Rather, these data exist in a world where other information pertaining to the same subject may be available. Through these additional sources of information, one may be able to reduce the anonymity of the anonymized dataset. This is referred to as a *linkage-attack* [33] and the dilemma is that one likely does not know what additional sources of information exist, or will be created.



© Grant McKenzie and Hongyu Zhang;
licensed under Creative Commons License CC-BY 4.0

12th International Conference on Geographic Information Science (GIScience 2023).

Editors: Roger Beecham, Jed A. Long, Dianna Smith, Qunshan Zhao, and Sarah Wise; Article No. 9; pp. 9:1–9:15

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Driven by the ubiquity of context-aware technologies, and the data they collect, we have seen a shift towards the development of computational approaches that leverage these data to model *places* [22, 26]. Through these approaches, photographs, audio recordings, temperature sensors, etc. are being used in combination with geographic data to provide more holistic representations of our environment and the places we inhabit. The irony is that the same data used to generate increasingly intricate models of the world can be used to violate one’s privacy and de-anonymize personal location information.

Within the privacy and anonymity domains, there have been considerable efforts on developing techniques that provide a trade-off between privacy preservation and data utility. Driven by the needs of individuals, most privacy models parameterize that trade-off, permitting users to exhibit control over their privacy based on their personal comfort levels. One of the most popular privacy preservation method for individual data sharing is *k-anonymity* [32]. The objective of this approach is to anonymize a data point such that it cannot be differentiated from $k-1$ other data points. Within geographic domains, these data points tend to be locations. For a wide variety of reasons (see [2]), an individual may want to obfuscate their location by reporting a set of locations (including their own), rather than their actual position alone. *Spatial k-anonymization* was introduced to address a number of challenges unique to geographic content [1, 7].

Much of the existing methodological work on *k-anonymity* and location privacy research is designed to be domain agnostic. Researchers overwhelmingly approach locations as simple geometric objects. In real-world scenarios, however, these objects represent entities that have a variety of properties and relationships. Furthermore, these entities do not exist solely in this dataset, i.e., other sources of related information exist. In this work, we explore such a real-world scenario and demonstrate how the privacy guarantee of a spatial *k-anonymized* dataset can be violated through the inclusion of external data. The real-world scenario of interest to us, is the process of sharing one’s location. This is a process that happens millions of times a day as people *check-in* to a location through social media, share their favorite restaurant with friends, or tag their location in a photograph. In these scenarios, *location* refers not to one’s geographic coordinates but rather the *place* that one is visiting, e.g., Mel’s Diner. The dilemma is in the trade-off between preserving privacy and sharing location data to gain utility. While I may be content to publicly share my visit to a trendy restaurant I may not wish to disclose the location of my teenager with anyone other than immediate family. It may still be useful, however, for my teenager to share anonymized location information, such as a set of k possible places, in order to receive recommendations for events nearby, for example.

The complexity of using places in a *k-anonymity* model is that an extraordinary amount of information is publicly available about most places, information that can be used to reduce the anonymity of someone sharing their platial location. In this work, we leverage the fact that different types of places have different visiting behavior and different hours of operation. For instance, people typically visit restaurants for lunch and dinner and more so on weekends than weekdays. The place types themselves also vary in popularity, regardless of time of day. For instance sports bars consistently receive more visitors than dentist offices.

While companies like Foursquare and Google collect the opening hours, popular visitation times, and overall popularity of most places in the world, access to this volume of data is unrealistic for most. For this work, we aggregate such data to the level of place type (e.g., Coffee shop) instead of place instance (e.g., Compass Coffee on 14th St.) and demonstrate that even a sample of place instances aggregated to this level can significantly reduce the anonymity of a place in a *k-anonymized* spatial dataset. More specifically, we will address the following three research questions (RQ).

- RQ1** Does the ability to identify an individual's location within a set of locations increase if we know the time the individual visited the location? Specifically, we investigate the degree to which temporal visitation patterns (signatures) can be used to reduce the efficacy of the spatial k -anonymity technique.
- RQ2** Do all temporal popularity signatures have an equal impact on the de-anonymization of a k -anonymized spatial dataset? We compare three types of temporal patterns and popularity values to identify which of them has the largest impact on the anonymity of an individual. We then determine if a weighted combination of these temporal popularity signatures can outperform the individual signatures.
- RQ3** Given a set of weighted temporal popularity signatures, by how much must we increase the number of places (k) in order to maintain the same level of anonymity promised by a non-enhanced k -anonymized spatial dataset? Furthermore, if a set of places are reported as a geographic region, what impact does the increase in k have on the average *size* of the reported region?

2 Related Work

A large body of literature pertaining to computational approaches to location privacy and anonymity has been published over the past few decades. Computational science research has mostly approached this from a geometric perspective [14, 16] whereas human geographers have typically taken a more qualitative approach [38, 12].

The concept of k -anonymity was first proposed by Sweeney and Samarati in 1998 [28] and later formalized as a property of certain anonymized datasets. *Relational k -anonymity* was then proposed as an approach for database privacy and disclosure control. A table is said to be k -anonymized if each record is indistinguishable from at least $k-1$ other records [32]. Within relational k -anonymity, *generalization* is often applied to reduce the uniqueness of each record, thus preserving a level of anonymity. While k -anonymity was originally designed with anonymity of the individual (or record) in-mind, an extension, ℓ -diversity [18], was proposed with the objective of preserving the *sensitivity of the values* associated with the records or individuals. This is addressed by introducing ℓ "well-represented" sensitive attribute values in each anonymized group. Li et al. [15] discovered that in some cases (e.g., skewed distributions or similar attributes) ℓ -diversity is insufficient in privacy protection. As a result, they propose t -closeness [15] to overcome the limitation of ℓ -diversity.

Spatial k -anonymity incorporates location information into the discussion of anonymity and privacy preservation. While relational k -anonymity is static and often involves a single k , the spatial version was designed to be dynamic with variable k [6].

Existing research on this topic has leveraged spatial k -anonymity for the development of *k -anonymized spatial regions* that include an anonymized set of locations consisting of an anonymized user and at least $k-1$ other users [6, 23, 8]. Early work by Kalnis et al. [11, 10] developed a series of cloaking techniques (e.g., Hilbert cloak, center cloak) with the goal of reducing vulnerabilities in basic spatial k -anonymity algorithms. Additional efforts have introduced techniques that consider the temporal connectivity of location-based services [5]. To date, the majority of research from computational scientists has approached spatial k -anonymity through the introduction of spatial-temporal cloaking and tree-based spatial indices, predominantly focusing on the geometric properties of the data.

Aside from *spatial k -anonymity*, additional methods of *geomasking* have been developed to obfuscate location information. While not strictly anonymity approaches, these are typically categorized into aggregation-based or perturbation-based with aggregation methods

being similar to anonymized spatial region’s *spatial-temporal cloaking* but often leverage existing geographic units such as administrative boundaries [3], Voroni polygons [30, 25], or census tracts [17]. Others have built aggregation techniques based on geometric shapes or centroids [36]. *Perturbation* geomasking methods displace individual data points to nearby locations using random distance and direction and various kernels [36, 37]. Finally, efforts have been made to combine the two types of geomasking methods. *Adaptive areal elimination* first aggregates population polygons into anonymized spatial regions, then randomly displaces data points within the newly formed anonymizing regions [13]. Charleux and Schofield [4] proposed *adaptive areal masking* that replaces longest border shares in adaptive aerial elimination with Euclidean distance ranks.

In recent years we have seen a substantial increase in computational approaches to define and understand *places*. As geographic information science has evolved, researchers are not only exploring the Aristotelian view of place (i.e., objects in the Euclidean space), but also the Platonic view (i.e., relationships and experiences in the environment) [27]. A growing body of work has been examining and modeling the concept of place from multiple dimensions [35, 29, 22]. As increased availability of large heterogeneous datasets from a variety of sensors has allowed geospatial scientists to move from spatial studies to the multidimensional concept of place, so too have the geoprivacy and spatial anonymity domains.

3 Data

3.1 Temporal visitation, hours of operation, and place popularity

Two different data sources were used in these analyses. First, all of the place types (e.g., Bars, Parks, Police Stations) published by the local place recommendation service, *Foursquare*, were identified.¹ We randomly selected 20 places of interest (POI) from across the United States in each of the place types. The Foursquare application programming interface (API)² was used to request the number of check-ins to each of these POI every hour over the course of 3 months. These check-in counts were grouped by place type and aggregated to hour of the week producing a set of 168 (24×7) temporal signatures (T_F) for each Foursquare place type. Hours of operation were accessed from the API for each of the Foursquare POI in our sample. These data consist of a binary value for each hour of a typical week. As before, these were grouped by place type and aggregated (median) to the hour of the week producing an hours of operation signature (T_H). Foursquare also offers a popularity value for each POI which is computed based on foot-traffic and user ratings.³ Using the Foursquare API, we accessed the popularity values for each POI in our sample dataset, and averaged them by place type. This produced a mean popularity value, Pop , for each place type in the Foursquare dataset.

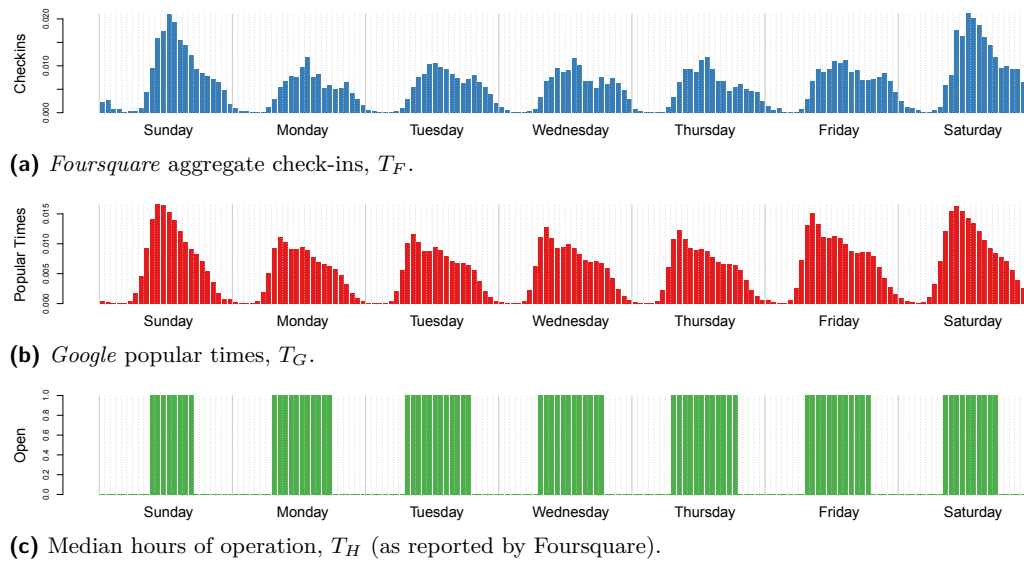
We then accessed popular times data for 185,600 *Google Places* POI across the United States.⁴ The popular times data are constructed through passive collection of location information accessed from the mobile devices of Google’s location service users. Similar to the process used for the Foursquare data, these popular times were groups by Google’s place

¹ A full list is available at <https://location.foursquare.com/places/docs/categories>.

² <https://developer.foursquare.com/>

³ <https://medium.com/foursquare-direct/tagged/engineering>

⁴ Data collection script available at https://github.com/apollojain/popular_times



■ **Figure 1** Example temporal signatures for the place type *Café*.

type⁵ (different from *Foursquare*'s) and aggregated by hour of the week. This approach produced a set of temporal signatures (T_G) for each *Google* place type. Finally, all three temporal signatures (T_F , T_H , T_G) were normalized individually producing a distribution of temporal values that sum to 1 (Figure 1). This normalization process was necessary so that each signature was evenly weighted at the start of analysis.

3.2 Place type alignment

Given the two sources of POI data, the first task was to align the place type schemas. We leveraged our previous work on this topic [20] to identify alignments between place types. The process involved collecting the same POI representations (e.g., the same restaurant) from *Foursquare* and *Google* via their APIs. POI matching was done by comparing the names and geographic distances between place representations. We took an overly conservative approach by only accepting matches for POI where there was an exact name match and the geographic distance was less than 100 meters. We then generated a matrix counting the occurrence of place type matches. The place types that had the largest number of POI matches were accepted as an alignment. For instance, *Foursquare* has a place type *Coffee Shop* while *Google* does not. Through our alignment process, we identified *Google*'s *Café* place type as a match. As a final step, we manually reviewed the alignment results and made minor adjustments to the place type alignments where appropriate.

3.3 Validation data

To validate our approach we required access to a large sample of data where an individual recorded their real-world visit to a location, including the time and place type they visited. While geosocial media *check-ins* are suitable for this task, access to a large and randomized

⁵ https://developers.google.com/maps/documentation/places/web-service/supported_types

sample is not possible directly through Foursquare or its gaming application, *Swarm*.⁶ Users of both Foursquare and Twitter, are able to connect their two accounts allowing them to publish their Foursquare check-ins on their public Twitter feed. Leveraging this knowledge, we used the public Twitter API⁷ to randomly sample 17,909,516 geotagged tweets within the continental United States between May 2017 and May 2022. The tweets were filtered to select only those whose source was Foursquare’s Swarm application. All of these tweets contained the information necessary to access a user’s geosocial check-in. Each check-in consists of the Foursquare POI name, place type, geographic coordinates, and timestamp of the visit. A total of 54,568 check-ins to 22,206 unique POI were identified after cleaning. To reduce POI bias we elected to only include one check-in (randomly selected) for each POI in our analysis.

Through the Foursquare API, we requested the closest set of POI to each of the 22,206 check-in POI. The API sets an upper limit of 50 POI per *Nearby* request. The maximum of 50 POI was not always returned, so in order to maintain a robust set of data, we removed all check-ins with fewer than 29 nearby POI from further analysis. This resulted in a final validation dataset of 19,478 check-ins to the same number of unique POI and a total of 584,340 nearby POI.

4 Analysis

Our first task in addressing RQ1 was to determine the degree to which one of our temporal signatures impacted our ability to identify an individual’s POI location from within a set of k POI. We will refer to an individual’s actual location as p_l , nearby locations as p_n , and the larger set of all 30 POI in a region as P_{30} , where $p \in P$. Each p has a *place type* and each visit to a p_l occurred at some time, reported as the hour of the week. Three temporal signatures and the popularity value were assigned to each p based on its place type.

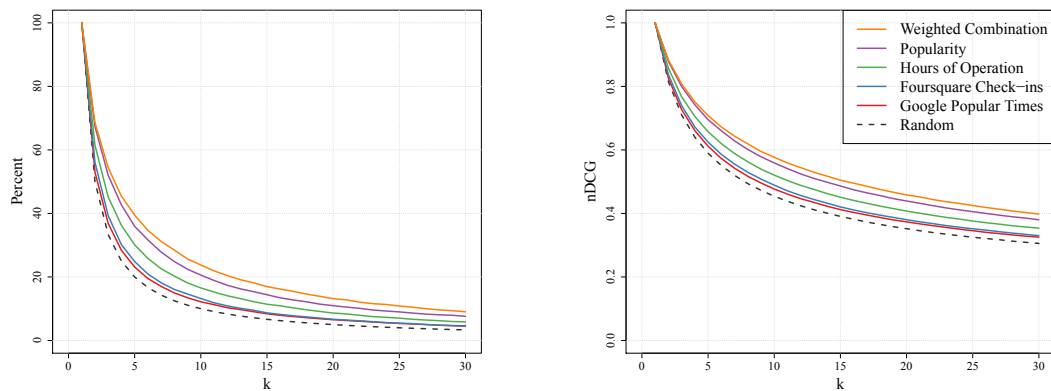
4.1 Spatial k -anonymization

We started with a baseline k -anonymized spatial dataset, one that includes p_l and a set of p_n nearby POI, but ignores the place type property or temporal signatures of each p . We set a range for k from 1 to 30 for our analysis. For each of the 19,478 check-in in our dataset, we selected a subset of the k closest POI (P_k) to p_l , including p_l itself. For instance, $k = 3$ means that P consisted of 3 p , including 2 p_n and our p_l . To determine the level of k -anonymity in our set, we randomly selected a p from the set of P_k . This was done for all 19,478 check-ins and all values of k . The average number of times p_l was correctly identified in P_k was recorded. The results are shown as the dashed black line in Figure 2a (the other lines will be discussed in Section 4.2).

Provided no other information on which to select a p from P_k , the results are random with the percentage equating to $1/k \times 100$. While informative, this method of only counting instances where p_l is correctly identified ignores position ranking. For example, a model that identifies p_l as the second most likely place is better than a model that identifies p_l as the 20th most likely place. This is irrelevant for the random model, but will play a role in assessing the temporal signature approaches. To account for differences in rank, we calculated the *normalized Discounted Cumulative Gain* (nDCG) (Equations 1 and 2). nDCG

⁶ <https://www.swarmapp.com/>

⁷ <https://developer.twitter.com/>



(a) Percentage of the time p_l is correctly identified in P_k .

(b) Normalized Discounted Cumulative Gain based on position of p_l in ranked P_k .

■ **Figure 2** Percentage of POI that were identified correctly, or where they ranked, using different approaches, shown as k increases.

considers the rank of a prediction by penalizing incorrect p_l selections at a \log_2 rate based on their position i in the ranking, where rel_i is the graded relevance of p in P_k . $IDCG_p$ is the idealized ranking where p_l is correctly identified in the first ranked position. The results of the nDCG assessment of the random spatial k -anonymity approach are shown in Figure 2b.

$$nDCG_p = \frac{DCG_p}{IDCG_p} \quad (1) \quad DCG_p = \sum_{i=1}^{|REL_p|} \frac{rel_i}{\log_2(i+1)} \quad (2)$$

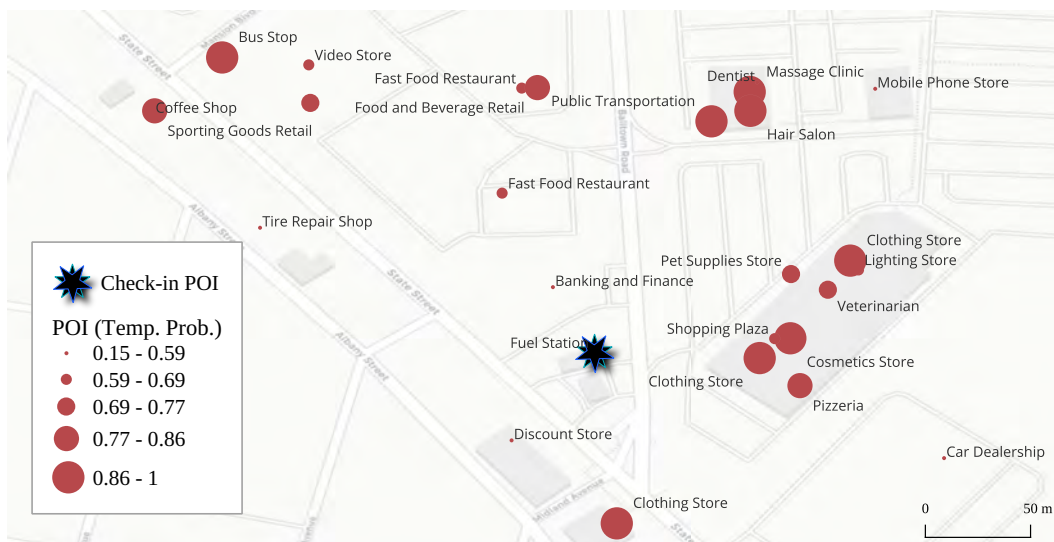
4.2 Temporal signature enhancement

We then designed a method to reduce the anonymity of a k -anonymized spatial dataset through the inclusion of place type temporal signatures. Our answer to RQ1 depends on whether the ability to identify someone increases with the inclusion of this temporal dimension.

To start, we limited our analysis to include temporal check-in behavior as reported by Foursquare at the place type level. Remember that each p in our dataset has a place type, and each place type has a Foursquare temporal signature, T_F . The check-in time for each of our p_l was recorded and used to identify the temporal probability of an individual visiting a p based on the temporal signature. For example, the temporal probability at 20:00 on a Friday is higher for the *Restaurant* place type than *Bank*. Figure 3 represents these temporal probabilities as graduate symbols.

As before, a subset of POI, P_k closest to p_l were selected. The p in this subset were then ranked based on the visitation probability at the indicated time (temporal signature). Given that P_k may contain multiple p of the same place type, these p have the same temporal probability value. Order was randomized between places of the same type. The p with the highest temporal probability was flagged as the predicted location of p_l . This was done for all known POI visits in our dataset and the average accuracy was reported for each value of k both as the *correctly* identified p and the nDCG. These are shown as the blue lines in Figure 2.

The results of this analysis indicate that the inclusion of place type temporal signatures increases one's ability to identify an individual's location in a set of k -anonymized POI. Averaged across all selected values of k (1-30), the inclusion of Foursquare's temporal signatures, T_F , decreased the anonymity of p_l in P_k by 31%. As shown in Table 1 (column T_F), the percentage of de-anonymization increases with larger values k .



■ **Figure 3** Places of interest in a region shown with graduating symbology representing the temporal probability at 20:00 on Friday. The black star marker indicates the actual location of the individual. Base map by Carto.

■ **Table 1** Average percentage improvement in correctly identify an individual (p_i), above random selection from a k -anonymized spatial dataset. The three different temporal signature-based approaches are reported along with the popularity value method and the weighted combination of temporal popularity signatures, $TPop$.

k	T_F (%)	T_G (%)	T_H (%)	Pop (%)	$TPop$ (%)
2	12.2	6.1	12.0	34.7	37.2
5	24.2	15.7	27.6	79.5	97.6
10	32.1	21.8	35.0	106.0	137.5
15	30.4	26.2	38.5	116.2	154.4
20	33.6	30.8	40.4	119.4	163.2
25	36.0	34.5	41.8	125.0	171.5
30	38.2	38.0	43.4	127.7	172.3

4.3 Comparing temporal signature and popularity approaches

Knowing that a place type temporal signature can be used to decrease the anonymity of an individual in a k -anonymized spatial dataset, we compared temporal signatures from different sources as well as the atemporal place type popularity values.

4.3.1 Temporal signatures

Having developed a model based on Foursquare’s temporal signature in the previous section, we conducted the same analysis for the Google popular times signatures T_G and the place type averaged hours of operation, T_H . As shown in Figure 2, ranking POI based on the probability of an individual visiting them at a given time improved the place prediction in all cases and for all values of k . In other words, location privacy was reduced through the inclusion of any temporal signature data. In comparing the results of analysis using different temporal signatures, T_G has the lowest impact, reporting an average decrease in anonymity of 26.3% across all values of k . Similar to T_F , the percentage increased with larger values of k . The T_H signatures produced the largest impact on anonymity with an average decrease of 35.3%. The percentage decrease in anonymity is shown for select values of k in Table 1.

4.3.2 Popularity

In addition to place type temporal signatures, the popularity of place types can also be used to reduce anonymity of a user's location in a set of POI. While the previous data signatures reported a relative change in visitation popularity over time, our popularity values, Pop , are atemporal and represents a comparison between place types, ranging from 0 (least popular) to 1 (most popular). These place type popularity values were assigned to their respective p in P_k and were ranked based on this popularity. We again randomly order p of the same place type within this ranking. As shown in Table 1, this approach results in a greater percentage of anonymity decrease than each of the temporal signatures alone. If we examine $k = 8$, for instance, there is a 1 in 8 (12.5%) chance of randomly selecting an individual's actual location in a spatial k -anonymized dataset. Through the inclusion of place type popularity, this doubles to 1 in 4 (25.0%). These results, along with those from the previous section address the first portion of RQ2, namely that all of these data signatures decrease anonymity by different amounts.

4.3.3 A weighted combination of signatures

In addition to assessing each of the temporal signatures and the popularity values independently, we also computed a weighted combination of the signatures. In addressing the second portion of RQ2, we question whether combining the signatures and popularity value will outperform, with respect to de-anonymization, each signature alone. The combined approach is shown in Equations 3 and 4. In our analysis, applied all combinations of weights, incrementing by 0.1 so that 285 combinations were applied to all temporal signatures and the popularity value. This was done for all 19,478 check-ins and all values of k between 1 and 30.

$$w_1(T_F) + w_2(T_G) + w_3(T_H) + w_4(Pop) \quad (3) \quad w_1 + w_2 + w_3 + w_4 = 1 \quad (4)$$

The results of this weighted approach, with all combinations of weights are provided in the project repository. The weight combination that produced the highest number of correct POI identifications, and highest nDCG, consisted of a weight of 0.3 for each of the temporal signatures and a weight of 0.1 for the average place type popularity. We refer to this weighted combination as the temporal popularity signature, $TPop$. On average, this approach decreased anonymity by 143.3% with exact values shown in Table 1. This is a substantial amount as compared to each of the temporal signatures and popularity independently.

We further investigated the results of this analysis by ordering all weighted combinations by their average accuracy across all values of k . Our top model of 0.3 for all temporal signatures and 0.1 for popularity values was ranked 1 out of 285 possible combinations. The first combination of weights to not include average place type popularity ($w_4 = 0$) was at rank 220. This suggest that the inclusion of popularity in our model is essential for a large decrease in anonymity, but that the actual weight is less important. Also of note, the best performing combination placed equal weight on each of the temporal signatures, indicating that each temporal signature represents a unique aspect of place visitation behavior and that all are needed in order to produce the best approach for de-anonymization of a k -anonymized dataset.

4.4 Platial k -anonymization

The results of the previous sections demonstrate that an attacker with access to temporal and/or popularity data reported at a place type level can considerably decrease the anonymity of an individual's reported location within a k -anonymized set of POI. The accessibility to,

and inclusion of, such contextual data requires that the number (k) of POI in a k -anonymized set be increased in order to guarantee the same level of anonymity promised by the original spatial k -anonymity model. In addressing RQ3, we establish these new values for k proposing that the values be labelled *Platial k* or k_p .

Through referencing the results of our analysis in Section 4.3, we can match accuracy percentages between a spatial k -anonymized (random selection) approach and our most accurate platial approach, *TPop*, taking the k value from our most accurate model as k_p . In other words, how many k_p are needed in order to guarantee the same level of anonymity that was promised by a k -anonymized dataset that assumed no temporal popularity data were available? Table 2 shows the value of k from a standard k -anonymized dataset along with the k_p values necessary to achieve the same level of anonymity using our temporal popularity signatures.

■ **Table 2** k number of POI along with the k_p number of POI needed to preserve k -anonymity given the temporal signatures, popularity values, or combination temporal popularity signature.

k	$k_p T_F$	$k_p T_G$	$k_p T_H$	$k_p Pop$	$k_p TPop$
2	3	3	3	4	4
5	7	6	7	11	13
10	14	13	14	23	29
15	21	20	22	>30	>30
20	28	28	29	>30	>30

For instance, in order to limit one’s exposure to a 20% chance of being randomly identified in a set of POI (the equivalent of a k -anonymity of 5), one would need to include 13 POI, or a k_p-1 of 12. As shown in Table 2, in some cases, the number of POI needed to preserve k_p -anonymity was greater than the 30 POI we had in each of our check-in sample sets. Using these results, we can report k_p as a function of k , namely $k_p = 2.54k + 0.04k^2 - 0.88$.

4.5 Reporting platial k -anonymity through geographic regions

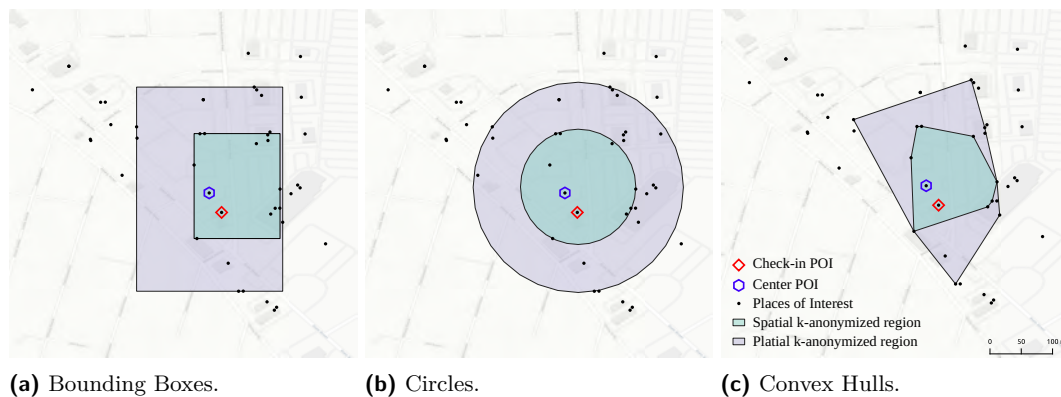
What do these results mean in practice though? The application of k -anonymity specifically deals with sets and spatial k -anonymity situates the elements of a set in geographic space. In real-world scenarios, anonymized spatial data are often reported through a location-based service as geographic regions, typically polygons that include the set of k -anonymized locations. Depending on the user’s privacy preferences, they set a large or small value for k which in turn determines the size of the reported polygon.

There are several ways to generate polygons that encompass a set of points. Here we identify geometric shapes based purely on the POI set, rather than political, social, or environmental boundaries. Such boundaries could also be used, but are not the focus of this work. The most common geometric shapes are a circle, bounding box, or convex hull. The centroid of these regions also varies. The simplest option is to set the centroid of the region on the known location and expand the radius or perimeter until k points are contained within the region. From an anonymity perspective, this approach falls victim to a *center-of-anonymized spatial region* attack, where an attacker would assume, given the geometry and centroid, that the actual location of an individual is the center most POI [11]. To avoid this, many current approaches [24] offset the centroid of the region by taking the n^{th} -nearest neighbor.

In generating a k -anonymized platial region, we have two options. One is *generalized* and involves simply referencing Table 2 or the k_p function to generate a polygon that contains k_p POI. This is a general approach as it uses the average k_p as reported through our analysis of

19,478 check-ins. While this can be used for any set of POI that contain place type attributes, platial k -anonymity can also be computed for an individual scenario. This is the *local* option. In this case, we assume the attacker has knowledge of the local region, knows the time someone visited a location, and has access to the place types of all POI. In this case, our set of P_k must include those that report a combined temporal popularity probability, $TPop$, greater than or equal to that of the actual check-in POI. The local platial k -anonymized region is the region that contains all of these POI. This may be better explained through an example. Let us set $k = 10$ and specify that our actual check-in POI has a $TPop$ probability value of 0.5. At a minimum, our platial k -anonymized region needs to include the 9 nearest POI with a $TPop$ probability greater than or equal to 0.5. Depending on the shape of the region, it may also include other POI with $TPop$ probability less than 0.5. All of these POI together sum to our local k_p .

Figure 4 shows examples of the various polygonal representations for a $k = 10$ anonymized set of POI as well as the k_p equivalent region. In these examples, the center point (blue hexagon), from which the shapes are determined, is the nearest neighbor to the actual *check-in* POI (red diamond).⁸ The geometries are generated by expanding the search radius from the center point until k POI are enclosed within the region. The smaller green regions show the minimum areas that encompass the specified k number of points (10 in this example), limited by the shape specifications. The larger purple regions represent the minimum areas that include P_k that are equal to or greater than the temporal popularity signature probability of the check-in POI at a given time. For a k of 10, k_p will always be at least 10.



■ **Figure 4** Polygonal representations of k -anonymity as well as the temporal popularity enhanced k -anonymity. These use a *local* approach based on the place types signatures of the actual POI.

For our sample set of 19,478 check-ins, we calculated the area of all three shapes that contain k and k_p POI. For all shapes and values of k , the areas of the platial k -anonymized regions are greater than the spatial k -anonymized regions. The difference in percentage decreases as k increases. The average percentage increase in area for k 1-20 ranges from 170.1% for a convex hull to 193.7% for a circle. Table 3 shows the median percentage increases in area. k is limited to 20 in this Table as we have seen that corresponding values of k_p can be considerably larger.

⁸ We use the first nearest neighbor here, but second or third could be used to increase privacy.

■ **Table 3** Median percentage increase in area between spatial k -anonymized regions and platial k -anonymized regions.

k	<i>Convex Hull</i> (%)	<i>Bounding Box</i> (%)	<i>Circle</i> (%)
2	0	520.9	845.7
5	394.0	343.6	324.0
10	105.9	96.9	120.9
15	73.9	67.4	105.5
20	32.7	37.8	56.8

5 Discussion

In the real-world, a geographic dataset does not exist in isolation. Additional information is available about all aspects of our lives, including the places that we visit. The times of day and days of the week that people interact with places in their environment follow patterns that can be discriminated at the categorical, or place type level. This knowledge can be leveraged and patterns can be used to estimate the locations of individuals. For the privacy-conscious among us, this is problematic. The spatial k -anonymity of a dataset states that an individual sharing a set of k places is guaranteed a level of anonymity.

In this work, we demonstrate that through the inclusion of place type temporal visitation patterns and popularity values, the presumed level of anonymity is violated. The results of RQ1 indicate that a place identification model built using publicly available temporal visitation signatures can significantly reduce the anonymity of a user sharing their location as a set of POI. Temporal signatures extracted from social media check-ins perform slightly better than those collected through passive data collection such as Google’s location services. Access to the average hours of operation for different place types outperform both of the activity-based temporal signatures. It is unclear exactly why hours of operation outperformed the behavior-based temporal signatures. One possible reason is that the hours of operation data were the least nuanced of the temporal data and by taking the median, the data were quite restrictive in reporting opening and closing times. It appears that for our sample of check-in data, these restrictive time periods were beneficial in predicting an individual’s location. By far the most useful information is the relative popularity of a place type. On average, access to these values substantially decreases the anonymity of an individual in a shared set of locations. This is worth noting as it suggests that the nuance of when a person visits a location, while important, is less important (on its own) than the overall, non-temporal popularity of a place. In identifying a weighted combination of these temporal signatures and popularity values (RQ2), we demonstrated that each of the different dimensions contributes to an improved model for de-anonymization. For instance, the probability of identifying an individual’s location out of a set of five POI ($k = 5$) is nearly 80% greater given access to popularity data and 100% greater using our weighted combination approach, compared to a model that did not include any additional data.

This equates to a meaningful decrease in individual privacy brought about by analysis of publicly available data. These signatures and popularity values are aggregated to the place type level, not the individual place instance, suggesting that they can be applied to k -anonymized POI datasets anywhere in the world. While research on temporal signatures has shown that roughly 50% of these temporal patterns vary regionally, some of the more common place types such as drug stores and restaurants, do not [19]. The results of our analyses demonstrate that k does not accurately represent the anonymity of a dataset given access to other sources of related data. To address this, we propose a *platial* value, k_p , that

represents the number of POI necessary to guarantee k -anonymity given an attacker may have access to these contextual data sources (RQ3). In this paper, we provide a reference for those developing place-based obfuscation applications, recommending a baseline k_p number necessary to ensure actual k -anonymity in a set of POI. Importantly, our proposed measure of k only assumes access to the three temporal signatures and one relative popularity set for a given set of place types. There are undoubtedly additional sources of information that can be used to further reduce the anonymity of a user sharing an anonymized dataset. In this work, we simply highlight some of the ways this can be done, and report the magnitudes of de-anonymization.

Our analysis reports that regions built from k_p -anonymized datasets are considerably larger in area than k -anonymized datasets. What was surprising was the dramatic increase in area reported on average. For instance, at $k = 5$, the average platial k -anonymized region was roughly 350% larger than the spatial k -anonymized region. Larger regions equate to a reduction in utility. While we argue that the anonymity of a user remains in-tact through our improved approach, the trade-off in utility must be acknowledged. All of this demonstrates a need for further critical investigation of how we choose to obfuscate location information.

The biases of the datasets used in this work must be mentioned. All of the data used in these analyses were contributed by individuals of geosocial media applications or a location service provider. While these data have been used in a wide variety of research, they do represent a biased subset of the population. Though check-ins were randomly sampled, the types of people that choose to check in and share their geographic locations are a unique subset of the population. They tend to be tech-savvy and predominantly live in urban areas. The data most often do not adequately reflect the activity patterns of the elderly, lower-income individuals, and those in rural communities. Any application or policy that uses the results of this work, should consider the biases and act accordingly.

A limitation of this work is the alignment of the two different place type vocabularies. Since Google and Foursquare use different terms and concepts to label their categories, alignment was necessary. As mentioned previously, the alignment was achieved through identifying co-occurrence of place instances. In some cases, a place type from one service would align with multiple place types from the other service. We took the place type that had the largest number of place instance matches, but sometimes the difference was a single POI. A manual check was done to ensure that the matches made sense, but any manual alignment introduces bias on the part of the person doing the aligning.

Future work in this area will involve the inclusion of additional contextual data such as the change in temporal behavior due to weather and local events. Our approach will be integrated with other efforts in the location privacy domain that leverage socio-economic, demographic, and mobility data. Additional efforts will be made in the application of this approach to real-world scenarios and privacy-preservation platforms, similar to projects such as *MaskMy.XYZ* [31] and *PrivyTo* [21].

6 Conclusion

In this paper, we identify some of the ways that the k -anonymity of an individual's reported location can be reduced by using existing publicly available place-based data. Specifically, our work shows that knowledge of place type temporal visitation patterns, average hours of operation, and relative popularity can substantially decrease the anonymity of one's location in a set of places of interest. Through analysis of 19,478 place check-ins we developed a platial k -anonymity approach that aims to improve anonymity, acknowledging that an attacker may have access to contextual information. Using this platial k -anonymized approach, we show that sets reported as geospatial regions must increase in area in order to preserve their

presumed degree of anonymity. Overall, this work demonstrates the need to be aware of the additional data that is increasingly available, publicly accessible, and can be used to reduce the anonymity of individuals sharing their seemingly obfuscated personal location information.

References

- 1 Charu C Aggarwal. On k -anonymity and the curse of dimensionality. In *Proceedings of the 31st International Conference on Very Large Data Bases*, pages 901–909, 2005.
- 2 Marc P Armstrong and Amy J Ruggles. Geographic information technologies and personal privacy. *Cartographica: The International Journal for Geographic Information and Geovisualization*, 40(4):63–73, 2005.
- 3 Marc P Armstrong, Gerard Rushton, and Dale L Zimmerman. Geographically masking health data to preserve confidentiality. *Statistics in medicine*, 18(5):497–525, 1999.
- 4 Laure Charleux and Katherine Schofield. True spatial k -anonymity: areal elimination vs. adaptive areal masking. *Cartography and Geographic Information Science*, 47(6):537–549, 2020.
- 5 Bugra Gedik and Ling Liu. Location privacy in mobile systems: A personalized anonymization model. In *25th IEEE International Conference on Distributed Computing Systems (ICDCS'05)*, pages 620–629. IEEE, 2005.
- 6 Gabriel Ghinita, Keliang Zhao, Dimitris Papadias, and Panos Kalnis. A reciprocal framework for spatial k -anonymity. *Information Systems*, 35(3):299–314, 2010.
- 7 Aris Gkoulalas-Divanis, Panos Kalnis, and Vassilios S Verykios. Providing k -anonymity in location based services. *ACM SIGKDD explorations newsletter*, 12(1):3–10, 2010.
- 8 Marco Gruteser and Dirk Grunwald. Anonymous usage of location-based services through spatial and temporal cloaking. In *Proceedings of the 1st international conference on Mobile systems, applications and services*, pages 31–42, 2003.
- 9 Alex Hern. New york taxi details can be extracted from anonymised data, researchers say. *The Guardian*, June 2014. (Accessed on 01/16/2023).
- 10 Panos Kalnis and Gabriel Ghinita. Spatial anonymity. In LING LIU and M. TAMER ÖZSU, editors, *Encyclopedia of Database Systems*, pages 2685–2690. Springer, 2009.
- 11 Panos Kalnis, Gabriel Ghinita, Kyriakos Mouratidis, and Dimitris Papadias. Preventing location-based identity inference in anonymous spatial queries. *IEEE transactions on knowledge and data engineering*, 19(12):1719–1733, 2007.
- 12 Carsten Kießler and Grant McKenzie. A geoprivacy manifesto. *Transactions in GIS*, 22(1):3–19, 2018.
- 13 Ourania Kounadi and Michael Leitner. Adaptive areal elimination (AAE): A transparent way of disclosing protected spatial datasets. *Computers, Environment and Urban Systems*, 57:59–67, 2016.
- 14 John Krumm. A survey of computational location privacy. *Personal and Ubiquitous Computing*, 13(6):391–399, 2009.
- 15 Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. t -closeness: Privacy beyond k -anonymity and l -diversity. In *2007 IEEE 23rd international conference on data engineering*, pages 106–115. IEEE, 2007.
- 16 Bo Liu, Wanlei Zhou, Tianqing Zhu, Longxiang Gao, and Yong Xiang. Location privacy and its applications: A systematic study. *IEEE access*, 6:17606–17624, 2018.
- 17 Yongmei Lu, Charles Yorke, and F Benjamin Zhan. Considering risk locations when defining perturbation zones for geomasking. *Cartographica: The International Journal for Geographic Information and Geovisualization*, 47(3):168–178, 2012.
- 18 Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkatasubramanian. l -diversity: Privacy beyond k -anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1):Article 3, 2007.

- 19 Grant McKenzie, Krzysztof Janowicz, Song Gao, and Li Gong. How where is when? On the regional variability and resolution of geosocial temporal signatures for points of interest. *Computers, Environment and Urban Systems*, 54:336–346, 2015.
- 20 Grant McKenzie, Krzysztof Janowicz, and Carsten Keßler. Uncovering spatiotemporal biases in place-based social sensing. *AGILE GIScience Series*, 1:14, 2020.
- 21 Grant McKenzie, Daniel Romm, Hongyu Zhang, and Mikael Brunila. Privyto: A privacy-preserving location-sharing platform. *Transactions in GIS*, 26(4):1703–1717, 2022.
- 22 Franz-Benjamin Mocnik. Putting geographical information science in place—towards theories of platial information and platial information systems. *Progress in Human Geography*, 46(3):798–828, 2022.
- 23 Mohamed F Mokbel, Chi-Yin Chow, and Walid G Aref. The new casper: Query processing for location services without compromising privacy. In *VLDB*, volume 6, pages 763–774, 2006.
- 24 Dilay Parmar and Udai Pratap Rao. Privacy-preserving enhanced dummy-generation technique for location-based services. *Concurrency and Computation: Practice and Experience*, 35(2):e7501, 2023.
- 25 Fiona Polzin and Ourania Kounadi. Adaptive Voronoi Masking: A Method to Protect Confidential Discrete Spatial Data. In Krzysztof Janowicz and Judith A. Versteegen, editors, *11th International Conference on Geographic Information Science (GIScience 2021) - Part II*, volume 208, pages 1–17, 2021.
- 26 Ross S Purves, Stephan Winter, and Werner Kuhn. Places in information science. *Journal of the Association for Information Science and Technology*, 70(11):1173–1182, 2019.
- 27 Stéphane Roche. Geographic information science ii: Less space, more places in smart cities. *Progress in Human Geography*, 40(4):565–573, 2016.
- 28 Pierangela Samarati and Latanya Sweeney. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. Technical report, Data Privacy Lab Report, 1998.
- 29 Simon Scheider and Krzysztof Janowicz. Place reference systems. *Applied Ontology*, 9(2):97–127, 2014.
- 30 Dara E Seidl, Gernot Paulus, Piotr Jankowski, and Melanie Regenfelder. Spatial obfuscation methods for privacy protection of household-level data. *Applied Geography*, 63:253–263, 2015.
- 31 David Swanlund, Nadine Schuurman, and Mariana Brussoni. MaskMy. XYZ: An easy-to-use tool for protecting geoprivacy using geographic masks. *Transactions in GIS*, 24(2):390–401, 2020.
- 32 Latanya Sweeney. k-anonymity: A model for protecting privacy. *International journal of uncertainty, fuzziness and knowledge-based systems*, 10(5):557–570, 2002.
- 33 Zhouxuan Teng and Wenliang Du. Comparisons of k-anonymization and randomization schemes under linking attacks. In *Sixth International Conference on Data Mining (ICDM'06)*, pages 1091–1096. IEEE, 2006.
- 34 J.K. Trotter. Public NYC taxicab database lets you see how celebrities tip, October 2014. (Accessed on 10/14/2022).
- 35 Daniel Wagner, Alexander Zipf, and Rene Westerholt. Place in the giscience community—an indicative and preliminary systematic literature review. In *Proceedings of the 2nd International Symposium on Platial Information Science (PLATIAL'19)*, pages 13–22. Zenodo, 2020.
- 36 Jue Wang, Junghwan Kim, and Mei-Po Kwan. An exploratory assessment of the effectiveness of geomasking methods on privacy protection and analytical accuracy for individual-level geospatial data. *Cartography and Geographic Information Science*, pages 1–22, 2022.
- 37 Paul A Zandbergen. Ensuring confidentiality of geocoded health data: assessing geographic masking strategies for individual-level data. *Advances in medicine*, 2014:1–14, 2014.
- 38 Hongyu Zhang and Grant McKenzie. Rehumanize geoprivacy: from disclosure control to human perception. *GeoJournal*, 88(1):189–208, 2022.

Data-Spatial Layouts for Grid Maps

Nathan van Beusekom  

TU Eindhoven, The Netherlands

Wouter Meulemans  

TU Eindhoven, The Netherlands

Bettina Speckmann  

TU Eindhoven, The Netherlands

Jo Wood  

City, University of London, UK

Abstract

Grid maps are a well-known technique to visualize data associated with spatial regions. A grid map assigns each region to a tile in a grid (often orthogonal or hexagonal) and then represents the associated data values within this tile. Good grid maps represent the underlying geographic space well: regions that are geographically close are close in the grid map and vice versa.

Though Tobler's law suggests that spatial proximity relates to data similarity, local variations may obscure clusters and patterns in the data. For example, there are often clear differences between urban centers and adjacent rural areas with respect to socio-economic indicators. To get a better view of the data distribution, we propose grid-map layouts that take data values into account and place regions with similar data into close proximity. In the limit, such a data layout is essentially a chart and loses all spatial meaning.

We present an algorithm to create hybrid layouts, allowing for trade-offs between data values and geographic space when assigning regions to tiles. Our algorithm also handles hierarchical grid maps and allows us to focus either on data or on geographic space on different levels of the hierarchy. Leveraging our algorithm we explore the design space of (hierarchical) grid maps with a hybrid layout and their semantics.

2012 ACM Subject Classification Theory of computation → Computational geometry

Keywords and phrases Grid map, algorithms, trade-offs

Digital Object Identifier 10.4230/LIPIcs.GIScience.2023.10

Supplementary Material *Software (Source Code)*: <https://github.com/nvbeusekom/dataspatial-hybridgridmaps>; archived at [swh:1:dir:49956cc7368207673acba37bc42be357ef4625f1](https://www.swh.io/dir/49956cc7368207673acba37bc42be357ef4625f1)

Acknowledgements We want to thank Kevin Verbeek for useful discussions on the computational aspects of this paper.

1 Introduction

Many types of data have a spatial component: they relate to regions of interests, points of measurement, countries or other administrative zones, et cetera. Visualizing such data in a geographic map then allows studying patterns in the data that may be influenced by local or global geography – such as observing differences between rural and urban areas, or between northern and southern municipalities. However, as the data complexity increases, visualizing all data in a standard, geographically accurate map becomes infeasible, as precise geography may cause objects to become indistinguishably small or to clutter. A solution is to warp the geography instead, to create a better canvas for portraying the data, understanding that precise geography is not a necessity for observing higher-level patterns.



© Nathan van Beusekom, Wouter Meulemans, Bettina Speckmann, and Jo Wood; licensed under Creative Commons License CC-BY 4.0

12th International Conference on Geographic Information Science (GIScience 2023).

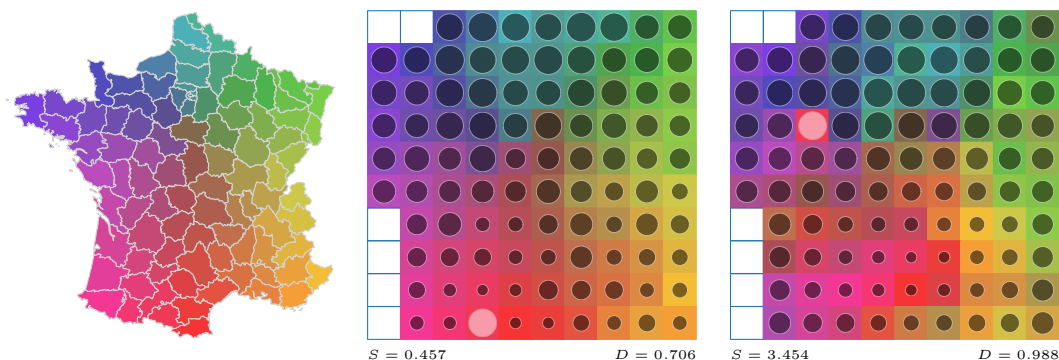
Editors: Roger Beecham, Jed A. Long, Dianna Smith, Qunshan Zhao, and Sarah Wise; Article No. 10; pp. 10:1–10:17

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

10:2 Data-Spatial Layouts for Grid Maps



■ **Figure 1** Synthetic example of France and two grid map layouts. Color indicates the region, circle size indicates the data value. The modified value is highlighted with a light circle.

A popular example of such a warped geography are *grid maps* (also called tile maps), which are used, for example, by news outlets [1, 2, 6, 8, 14, 15, 21] and in the geo-visualization literature [10, 17, 19, 25, 28]. Grid maps schematize each region in the input into a simple shape (a *tile*), often a rectangle or hexagon, to then subsequently arrange these tiles into a grid, roughly according to geography. These tiles then act as a container for a visual encoding of the data associated with each region, which can take as simple a form as coloring, but might be as complex as charts for multivariate data. Effectively, a grid map is a spatially conditioned and arranged small-multiples visualization [22].

We call the assignment of regions to tiles in a grid a *layout*. The layout of a grid map is traditionally determined by the geography of the regions. There are many situations where one can expect that the geography and patterns in the data correlate, at least to some degree. Tobler, in his “first law of geography” [20], expressed this expectation as follows: “everything is related to everything else, but near things are more related than distant things”. Local variations, however, may obscure interesting patterns in the data. For example, urban centers and adjacent rural areas frequently exhibit differences on socio-economic indicators, while cities, even when far removed from each other, often exhibit similar data values. Here, a purely spatial layout makes it difficult to obtain good overview of the data distribution. Given that a grid map already distorts space to some degree, one might hence consider layouts which distort space more to allow for a better representation of the data.

Consider the synthetic example of the departments of France in Figure 1. As data we are using the latitude of the region centroid (with some noise), with the exception of Ariège in the south, to which we assigned a northern value. In the grid maps color links cells to regions, and the circle size represents the data, that is, the latitude. The modified value is highlighted via a light circle. The middle of the figure shows a spatial assignment of departments to cells; regions are placed in a geographically coherent way. The right of the figure shows a *hybrid* layout that takes both space and data into account; regions are mostly placed to be geographically coherent, but also in such a way that similar values group together. The outlier can now easily be compared to regions with similar values, such that its place in the data distribution is clear. We posit that such hybrid layouts, that encode both data and space via location in the grid, might make it easier to observe patterns in the data that may otherwise stay hidden. Please note that, although we use color to indicate the effectiveness of our results, color is not required to indicate location in an eventual visualization and can hence be used to encode additional information.

Results and organization. In this paper we initiate the (algorithmic) study of hybrid layouts for grid maps. After covering some preliminaries in Section 2, we discuss in Section 3 measurements of how well space and data are represented in a given layout. Using our measures, in Section 4 we develop an algorithm that allows us to gradually transition from a spatial layout (which ignores data) to a data layout (which ignores space) and vice versa. Finally, in Section 5 we present a second algorithm which uses hierarchical properties of the geography to create hierarchical hybrids of spatial and data layouts. The resulting grid maps are visually pleasing and illustrate the potential of our techniques. We close in Section 6 with an extensive discussion of our results.

Related work. Algorithmically, grid maps were studied by Eppstein et al. [9], in the case that the number of tiles is roughly equal to the number of regions. The authors identified three main quality aspects: location (position in the grid with respect to the original geographic position in the map), adjacency (adjacent regions should map to adjacent tiles) and relative orientation (compass directions between regions). Eppstein et al. established that optimizing location is effectively a point-set matching problem, in which the sum of squared distances between the region centroids and their assigned tile centroids (the *displacement*) is to be minimized. Minimizing this displacement generally performs well also for the other criteria.

Meulemans et al. [11] broadened the perspective on grid maps by considering a suite of measures to assess layout quality in cases where the number of tiles well exceeds the number of regions. Here, empty tiles (gaps) can be used to give a more accurate reflection of the underlying geography. Generally speaking, a larger number of tiles allows for more accurate geographic representation at the expense of the size of the individual tiles, which in turn limits the complexity of the data visualization within each tile. The authors concluded that minimizing displacement still leads to good grid maps, but only in specific simple geographic settings. Subsequently, Meulemans et al. [12] described a pipeline to create high-quality grid maps for general geographic settings, by partitioning the input into simple pieces, leveraging cartogram techniques (see below) to create a tile configuration, and then minimizing displacement within each piece.

Grid maps have been applied hierarchically, for example, to create origin-destination maps (OD-maps) [18, 27]. OD-maps show the same (single) level both as the higher level structure, as well as the content of each tile, to visualize relational data between regions (such as migration). Here the hierarchical layouts are purely spatial, but they suggest nevertheless that nested layouts can be useful when investigating complex data.

Cartograms are another technique to overcome the constraints enforced by geographic detail. They deform the map such that every region has an area proportional to their data value. There are a wide variety of cartograms; most closely related to grid maps are contiguous cartograms that use schematic outlines, such as rectangular [4, 24], rectilinear [7], and mosaic cartograms [5]. The latter represent regions by multiple tiles in a grid, corresponding to data values; this is in contrast to grid maps which represent each region with a single tile. The quality of a cartogram is determined by two criteria that measure spatial coherence and data representation: (1) how well are the adjacencies and the relative positions of the geographic regions preserved in the cartogram, and (2) how large is the cartographic error (how well do the region sizes match their data values). Cartograms explicitly encode data, but similar data values cannot cluster unless the cluster is already present in the geography.

Treemaps show a hierarchy, using recursively partitioned rectangles, sized according to data values. Originally, there was no geographic space associated with the data, and optimization focused purely on achieving rectangles with low aspect ratios (i.e., close to

squares); see for example [3]. Ordered treemaps [16] provide more control over the treemap structure, by relating it to a one-dimensional ordering of the data elements. Wood and Dykes [26] proposed to use two-dimensional (geographic) space to control the treemap structure, resulting in spatial treemaps – effectively a hybrid of cartograms and treemaps.

2 Preliminaries

Our input is a geographic map which consists of a set $\mathcal{R} = \{r_1, \dots, r_n\}$ of n regions. Each region r_i has a polygonal representation p_i , a centroid c_i derived from it, and a data value v_i . Two parameters specify the target grid: its width W (the number of columns) and its height H (the number of rows). Together, they define a set $\mathcal{T} = \{t_{i,j} \mid 1 \leq i \leq W \text{ and } 1 \leq j \leq H\}$ of $W \cdot H$ tiles. Each tile is identified with its centroid. We focus on grid maps that use few tiles to represent the input regions, resulting in few gaps. That is, we set H and W such that $H \cdot W \geq n$, $(H - 1)W < n$ and $H(W - 1) < n$. Furthermore, we assume that the tiles in the grid and the geographic map have been aligned; see Eppstein et al. [9] for algorithms to optimize such alignment. We use square grids throughout our exposition, but our techniques readily generalize to other regular tile shapes.

The mapping of regions into a grid map is an injective function $L: \mathcal{R} \rightarrow \mathcal{T}$, the *layout*, which assigns each region to a unique tile in the grid. We use \mathcal{A}_L to denote the set of all (unordered) pairs of regions that are assigned to adjacent tiles in layout L .

Datasets. To illustrate our techniques, we use three different geographic maps, and population data per region for each:

FR: The 94 departments of continental France. Source: https://en.wikipedia.org/wiki/List_of_French_departments_by_population

EW: The 331 Lower Tier Local Authorities of England and Wales, hierarchically aggregated into Wales and the 9 regions of England. Source: <https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/articles/demographyandmigrationdatacontent/2022-11-02>

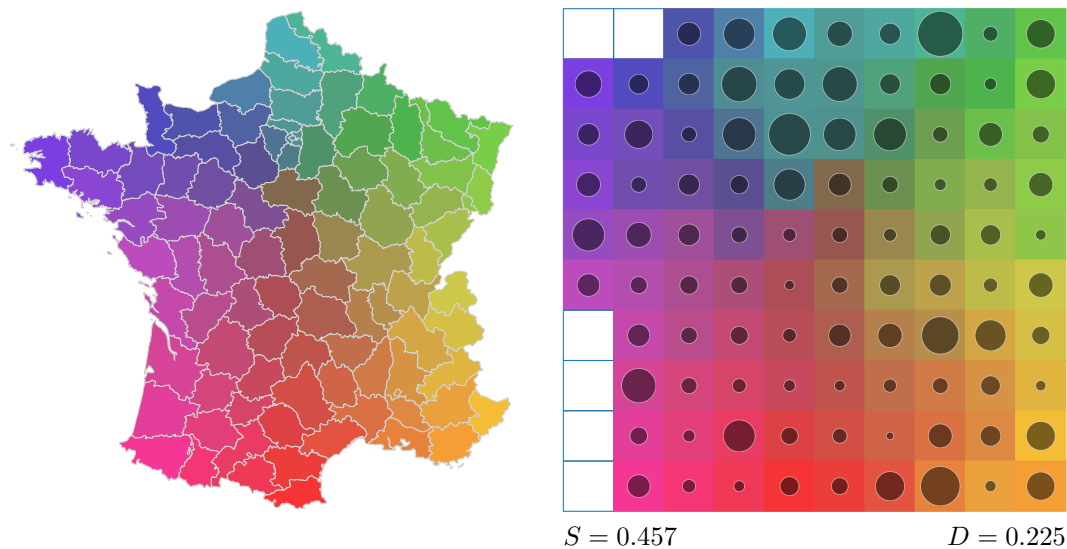
NL: The 388 municipalities of the Netherlands in 2017, hierarchically aggregated into 12 provinces. Source: <https://opendata.cbs.nl/statline/#/CBS/nl/dataset/70072ned/table?dl=4675>

3 Layout quality

To understand and measure the quality of hybrid layouts, which take both space and data into account, we need to be able to measure how well space and data are represented in a given layout. In the following, we hence consider the two extremes: spatial layouts that ignore the data values and data layouts that ignore the geographic space.

3.1 Spatial layouts

A *spatial layout* is based purely on the geography of the input map and ignores the data values; in other words, it is a traditional grid map. We hence use the results of Eppstein [9] to compute spatial layouts by minimizing the total squared displacement between the region centroids and the centers of the grid tiles. This method crucially relies on the alignment of the geographic map and the grid: translation of the map can result in a different layout. As mentioned in Section 2, we assume that the input map and the grid have been aligned well, using known methods [9]. The result on the France dataset is shown in Figure 2. In the following we discuss how best to measure how “spatial” a given layout is.



■ **Figure 2** The spatial layout for FR: color encodes location, circles sizes encode data values.

Spatial distortion and spatial correlation. We compute spatial layouts by minimizing displacement. Hence, a priori, displacement seems the logical choice to measure how well space is represented in a layout. However, displacement is inherently a global measure and as such not well suited for the local changes to the layout that occur during the kind of gradual transitions between spatial and data layouts that we envision for our hybrid layouts. *Spatial distortion* is a local measure for the distance and direction between regions that have been assigned to adjacent tiles. Specifically, with some normalization, we measure the spatial distortion S of a layout L as follows:

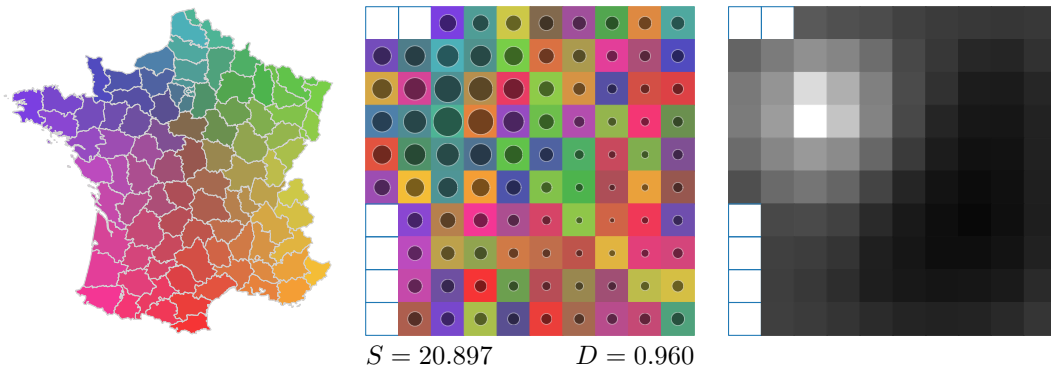
$$S(L) = \frac{1}{\mathcal{A}_L} \sum_{\{r_i, r_j\} \in \mathcal{A}_L} ((c_i - c_j) - (L(r_i) - L(r_j)))^2.$$

Here c_x and $L(r_x)$ are vectors expressed in unit lengths (i.e. tile widths). That is, we average the difference of the vectors between two adjacent tile centers and their assigned region centers. We use vector subtraction such that direction is also taken into account. We say that a layout L has high spatial correlation, when the spatial distortion $S(L)$ is low. Previous work [11] has established that minimizing displacement also generally results in low spatial distortion for spatial grid maps. We expect that minimizing $S(L)$ is NP-Hard: a reduction from Euclidean TSP may follow to minimizing $S(L)$ on a $n \times 1$ grid. Hence, our results are generally not (Pareto-)optimal.

In our figures, we color regions by their position in geographic space using a gradient. As a result, small geographic distances between adjacent tiles translate to low color variations – a layout that has high spatial correlation hence visually demonstrates smooth color changes.

3.2 Data layouts

A *data layout* is based purely on the data and ignores the geography of the input map. In contrast to a spatial layout – which corresponds to a traditional grid map – there is no one established way to create such layouts. We generally expect a data layout to cluster similar values together, especially at the high and low extremes of the value range. In a one-dimensional grid (that is, an array), this is readily achieved by sorting the regions by



■ **Figure 3** A data layout with a peak of high values, a valley of low values, and average values along the edges and through the diagonal.

value. However, in a two-dimensional grid, the optimum is less obvious. Before discussing our method to compute a data layout, we first define a measure for how well the data is represented in a given layout.

Data correlation. A grid map is by definition a map with well-defined edge adjacencies between tiles. Hence we can leverage spatial auto-correlation measures on the grid map space to measure similarities between data values of those regions that are mapped to tiles which are adjacent (or generally close) in the grid. Moran’s I [13] is a general measure for spatial auto-correlation which can be flexibly configured for various models of spatial proximity, via weights defined between each pair of regions. We want to focus on local relations and hence use a weight of 1 between adjacent tiles, and 0 between all other pairs. We defined the *data correlation* D of a layout L as the value of Moran’s I for a layout L :

$$D(L) = \frac{n}{|\mathcal{A}_L|} \frac{\sum_{\{r_i, r_j\} \in \mathcal{A}_L} (v_i - \bar{v})(v_j - \bar{v})}{\sum_{r_i \in \mathcal{R}} (v_i - \bar{v})^2}.$$

Here \bar{v} denotes the average of all v_i . The value of D is always between -1 and 1 , where values towards 1 indicate a strong correlation of data values, and values towards -1 indicate a strong inverse correlation of data values.

Regions whose data values are significantly higher or lower than the average contribute most to the data correlation if they are placed next to other regions with the same deviation. That is, in a layout with high data correlation, very low and very high values cluster together, as we would expect from a data layout. Note that tiles along the boundary have fewer neighbors than interior tiles and hence have a smaller impact on the data correlation. This naturally attracts tiles with average data values to the boundary and onto a diagonal, separating a peak and a valley, see Figure 3.

Computing layouts. As mentioned above, computing an optimal data layout in one dimension is simply sorting. So one can wonder if there is a two-dimensional equivalent? A naive approach would fill row after row by the next highest value – that is, effectively sort in the same manner as squarified treemaps do [3]. In such a layout many tiles can be expected to not be adjacent to other tiles of similar values in more than two of the four directions and hence the data correlation is generally low. There are many other possibilities to map a one-dimensional order of the regions onto the grid along a space-filling curve. While some

such mappings do better than others in terms of clustering similar data values, they all necessarily contain adjacent tiles with data values that are far removed from each other in the order. As a result, none of these sorted layouts optimize data correlation.

We hence computed (near-)optimal data layouts using simulated annealing, repeated for a set of random layouts. Effectively, we used our constrained annealing approach described in the next section, without constraints. We observed very slight variations in results due to randomization, but all with similar values for D . The layout with the highest data correlation, depicted in Figure 3, was used as the data layout in our experiments.

4 Hybrid layouts

In the previous section we introduced the two layouts that form the end of our spectrum: the spatial layout which ignores the data values and the data layout which ignores the geographic space. We also defined two measures to assess how “spatial” or how “data” a given layout is: the spatial distortion and the data correlation. In this section we now aim to compute meaningful *hybrid layouts* that represent both space and data to varying degrees.

Both measures are based on distances between tiles, but nevertheless, they do not live in the same mathematical space. It is hence unclear how to combine them in a meaningful way when computing and assessing hybrid layouts. We therefore take the following approach: we optimize for one measure while constraining the other. That is, we start on either end of the spectrum, constraining either space or data, and then optimize for the other, while slowly releasing the constraints. This creates two ranges of layouts, from space to data and vice versa, which allow us to explore the space of hybrid layouts and the corresponding trade-offs.

Simulated annealing. We use a constrained variant of simulated annealing: we define some slack parameter σ and we allow the algorithm to search only the space of layouts that are at most σ different from either the spatial or the data layout. Concretely, our algorithm is given an initial layout and performs random swaps of tiles. This follows standard simulated-annealing practice, with the addition that any swap that results in a layout exceeding the given slack σ from the initial layout is always rejected. We use a starting temperature $T_s = \frac{\delta_s}{\ln(0.5)}$, an ending temperature $T_e = \frac{\delta_e}{\ln(10^{-9})}$, 10^7 iterations, and exponential multiplicative cooling with factor $(T_e/T_s)^{10^{-7}}$. For improving the data correlation, we set $\delta_s = 10^{-3}$ and $\delta_e = 10^{-1}$. For reducing the spatial distortion we set $\delta_s = 10^{-2}$ and $\delta_e = 1$. Due to locality of our measures, each iteration takes $O(1)$ time.

Spatial to data. To explore the trade-off from the spatial extreme, we initialize the annealing algorithm with the spatial layout L_s and optimize for data correlation D , using a spatial slack parameter σ_S . Any swap that results in a layout L with $S(L) > S(L_s) + \sigma_S$ is rejected.

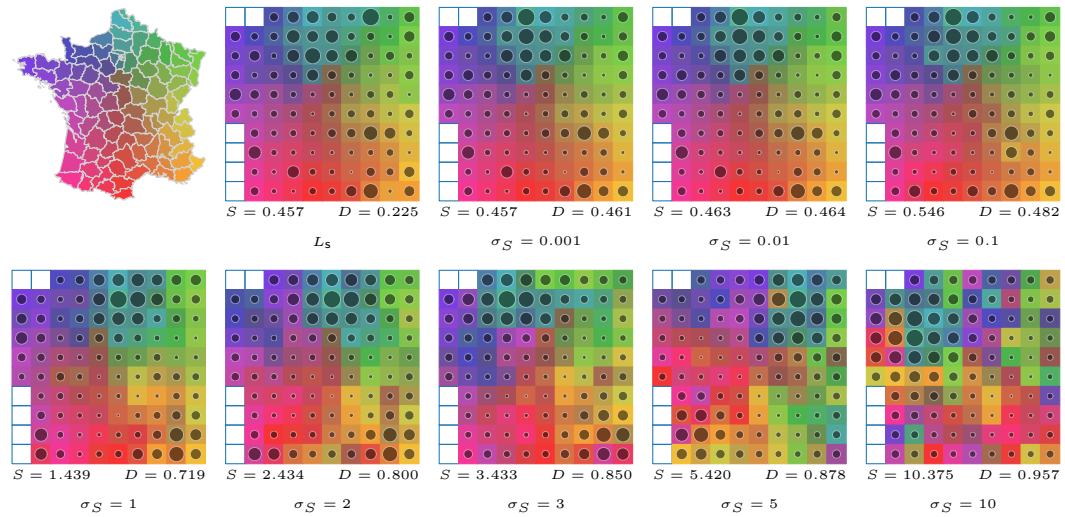
The results for France with increasing values of σ_S are shown in Figure 4. We observe that even for a tiny amount of spatial slack $\sigma_S = 0.001$, the data correlation already increases considerably. Apart from the measured improvement, we indeed observe that similar values are grouped together: a peak of high values occurs in the northern blue area. However, we do not observe further significant improvement until σ_S is increased to 1. For $\sigma_S = 1$ and $\sigma_S = 2$ two southern peaks emerge in the purple and orange area. These peaks are merged together into a single southern peak for $\sigma_S = 5$, yet still separated from the northern high values. At this point we also see the small values cluster in the center of the grid. Finally, the northern and southern peaks merge at $\sigma_S = 10$. This corresponds to an increase in data correlation D , which is already close to $D(L_d)$. At this point most of the spatial correlation is

10:8 Data-Spatial Layouts for Grid Maps

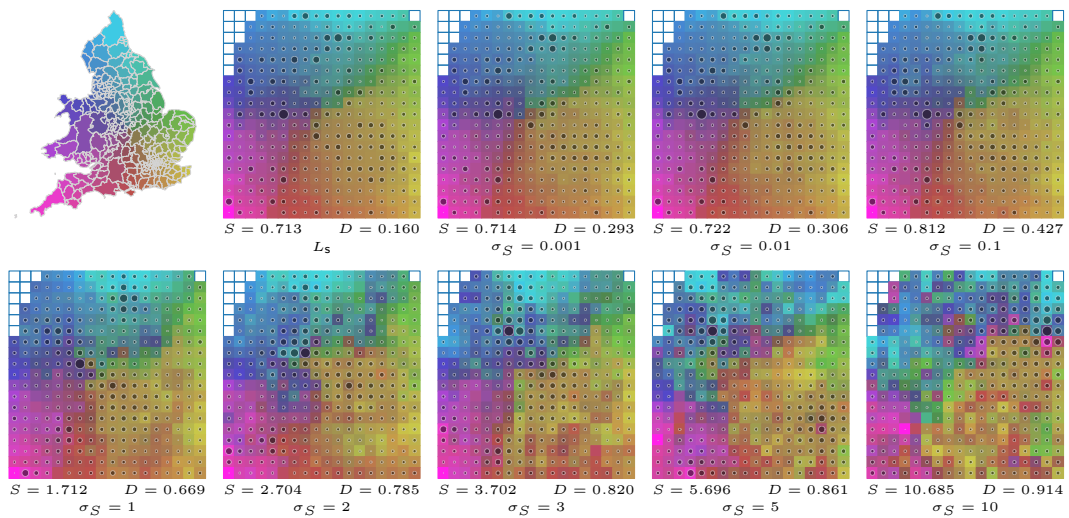
already lost. Increasing the spatial slack further leads to only minor improvement of D , while even further increasing the spatial distortion. We observe similar behavior in the results for EW (Figure 5) and for NL (6). However, the improvement of the data correlation happens more gradually. This might be due to the large number of regions, presenting the simulated annealing algorithm with more possibilities, while maintaining a similar spatial slack.

Data to spatial. To explore the trade-off from the data extreme, we use an analogous implementation of our constrained annealing algorithm. We initialize it with the data layout L_d and reduce the spatial distortion S , using a data slack parameter σ_D : any swap that results in a layout L for which $D(L) < D(L_d) - \sigma_D$ is rejected.

Figure 7 shows our results for France with increasing values of σ_D . We again observe that allowing a small data slack $\sigma_D = 0.001$ already leads to a considerable reduction of spatial distortion. Visually, the data distribution is nearly identical to L_d , yet regions are starting



■ **Figure 4** Hybrid layouts for FR with increasing spatial slack σ_S .



■ **Figure 5** Hybrid layouts for EW with increasing spatial slack σ_S .

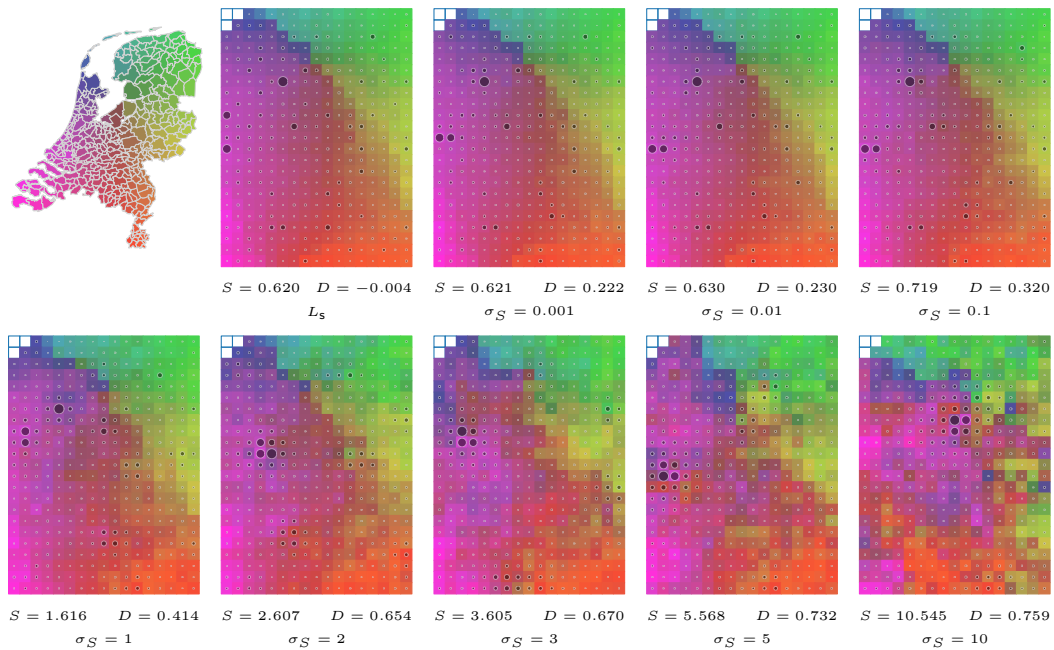


Figure 6 Hybrid layouts for NL with increasing spatial slack σ_S .

to cluster if they are geographically nearby, indicated by smoother color changes. The effect further strengthens for $\sigma_D = 0.005$, where it becomes easier to locate most regions with low and average values from some particular areas. As σ_D grows, more patches of similar colors appear, and less similar data values become adjacent. At $\sigma_D = 0.1$ the average values seem visually unorganized, and at $\sigma_D = 0.3$ some of the larger values are far from other peaks of large values. Most regions are now geographically grouped, apart from the orange regions. At $\sigma_D = 0.5$, most data correlation is lost, but the spatial distortion is close to $S(L_S)$. Again, the results for EW (Figure 8) and for NL (9) show similar behavior, though the improvement with low amounts of slack seems more significant. This might be due to

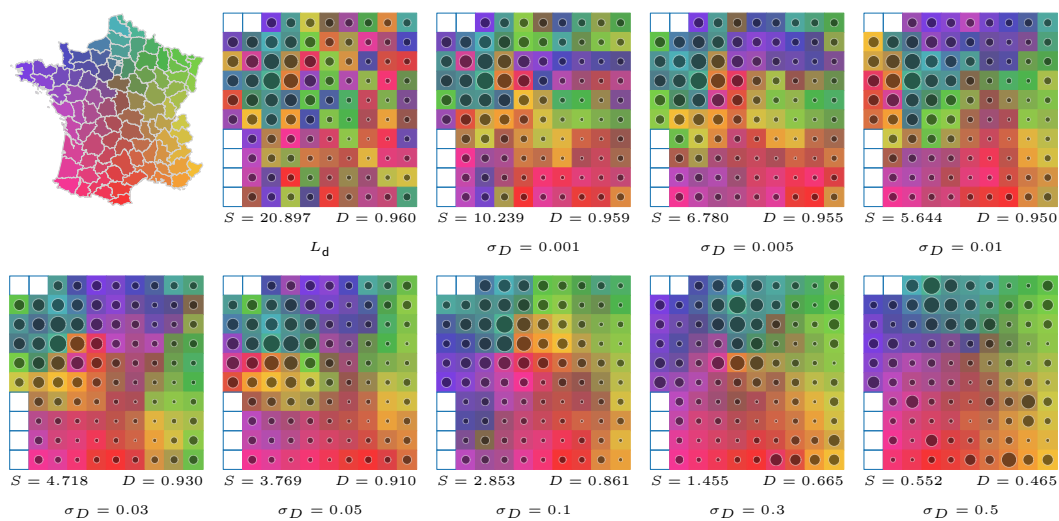


Figure 7 Hybrid layouts for FR with increasing data slack σ_D .

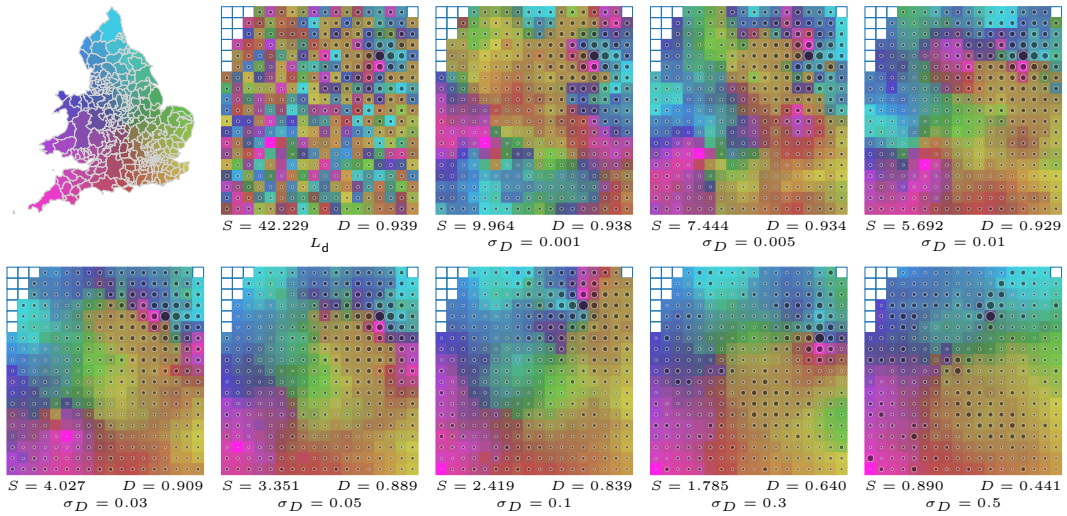


Figure 8 Hybrid layouts for EW with increasing data slack σ_D .

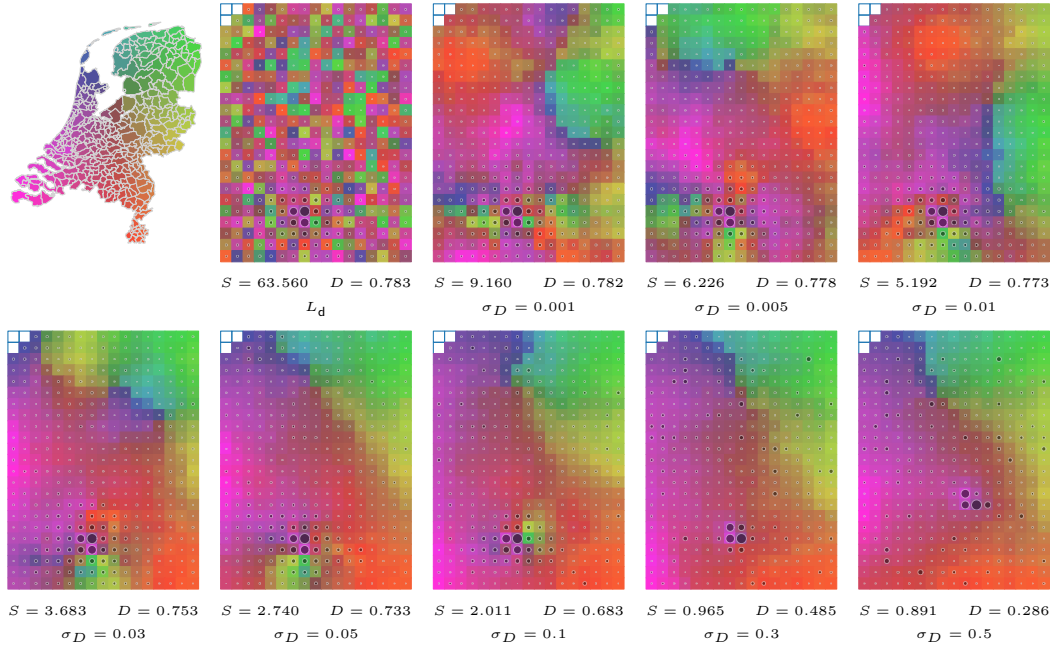


Figure 9 Hybrid layouts for NL with increasing data slack σ_D .

the many similar data values in EW regions, giving the algorithm many valid swapping possibilities. We observe that it takes a large amount of slack before the outlying values move to a more suitable spatial position, with the largest value not even being in an optimal spatial position at $\sigma_D = 0.5$.

Analysis. There are layouts in both sequences that strictly outperform layouts in the other sequence. For example, the layout of France with $\sigma_D = 0.001$ has higher data correlation *and* lower spatial distortion than the layout with $\sigma_S = 10$. Similarly, the layout with $\sigma_S = 1$ outperforms the layout with $\sigma_D = 0.3$. We conclude that the trade-off can be efficaciously approached from both ends using our constrained simulated-annealing approach. Both

methods readily achieve improvements using only small slack values. High slack values achieve reasonable results but are generally outperformed by the opposite approach with low slack values. We observe that the EW and NL layouts produced with high slack values are less close to the opposite extremes than is the case for France. They are hence still more of a hybrid layout. Increasing the slack further would eventually lead to the opposite extreme.

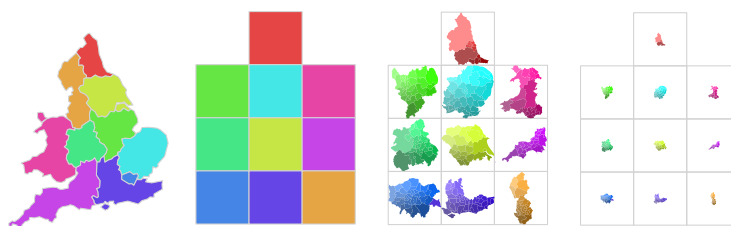
5 Hierarchical hybrid layouts

In the previous section we explored the trade-off between data and spatial layouts by creating sequences of layouts from both extremes. Each sequence focuses on one aspect, space or data, and integrates the other. While this approach arguably creates meaningful hybrid grid maps in the vicinity of the two ends of the spectrum, in the conceptual middle of the trade-off, the results might become difficult to interpret: for any given cluster of similar regions, it is unclear if this cluster was formed based on spatial or data correlation, which may hinder understanding of relations between clusters. Hence, in this section, we explore a more structured approach to integrate space and data. Specifically, we propose to use hierarchical information about the regions, effectively creating a *hierarchical grid map*. These zones may be administrative zones or deduced from the geography. It is less clear what meaningful hierarchies based on data would be; we hence focus on geographic hierarchies.

Within a zone regions may be arranged to create a spatial or data layout, while still displaying the zone affiliation. Meanwhile, the zones may also be arranged in a spatial or in a data layout, as to convey information about their relations on the higher level. For example, in a dataset with neighborhoods as regions and cities and rural areas for zones, the cities could re-arrange themselves to cluster together, separating themselves visually from rural zones. Yet, within each city (or within the rural areas), the spatial structure of neighborhoods is maintained. To achieve legibility across hierarchical level, it may be desirable to retain connectivity of regions within the same zone.

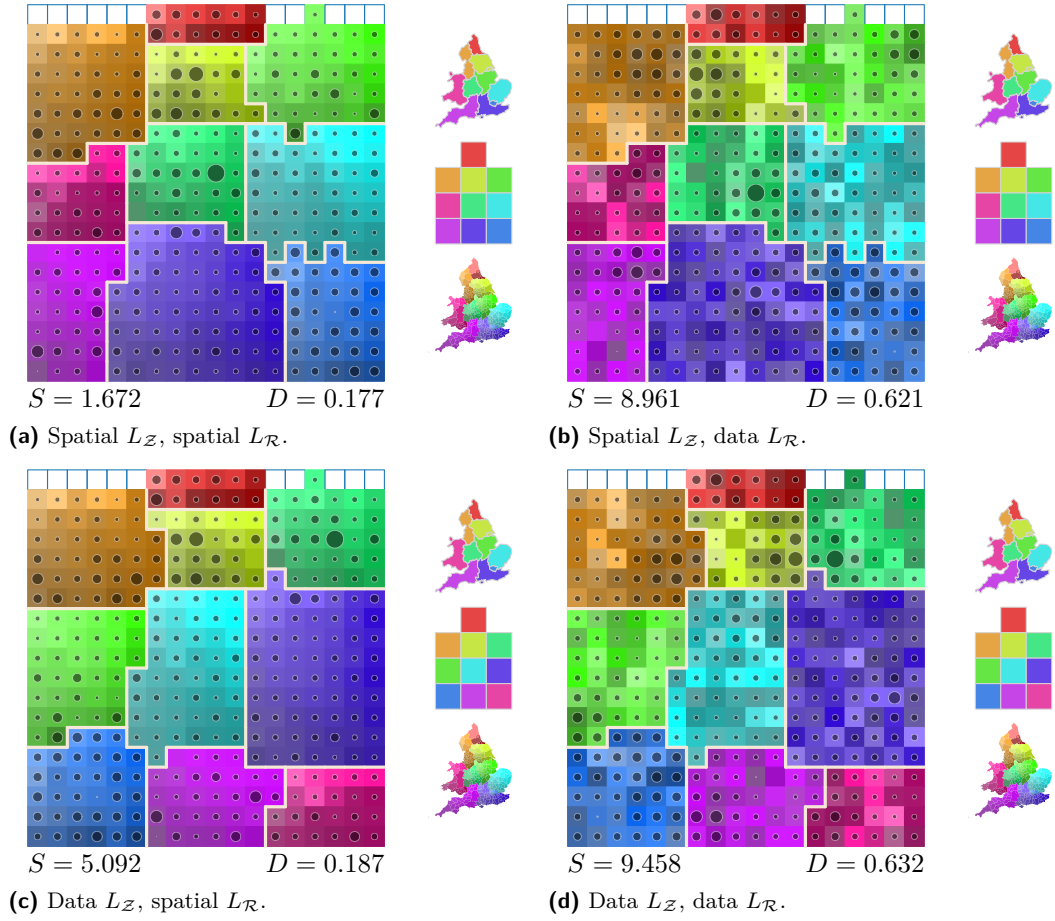
In the following we assume that our input map is augmented with a set $\mathcal{Z} = \{z_1, \dots, z_k\}$ of k zones. These zones partition \mathcal{R} , and we assume that they each capture a geographically somewhat coherent set of regions, such as provinces in a country. Furthermore, the data values of all regions in a zone z_i are aggregated into a value a_i for the zone – the form of this aggregation (e.g., average or sum) depends on the nature of the data and the desired effect in the map. We use administrative, established hierarchies in our experiments, aggregating based on the sum, as our data values represent population.

The central idea of our algorithm is then to leverage the hierarchy: we separate spatial and data aspects on different hierarchical levels, and finally blend the two together into a single layout. Each level of the hierarchy can independently be assigned to be a spatial layout, or a data layout. For simplicity, we assume a 2-level hierarchy. We hence select either data or spatial layouts on the zone-level and either data or spatial layouts on the region-level.



■ **Figure 10** Zone-level transformations: zones of EW; zone-level layout $L_{\mathcal{Z}}$ (data layout); regions of each zone fitted to the assigned tile; each zone scaled by factor λ , $\lambda = 0.25$ for illustration.

10:12 Data-Spatial Layouts for Grid Maps

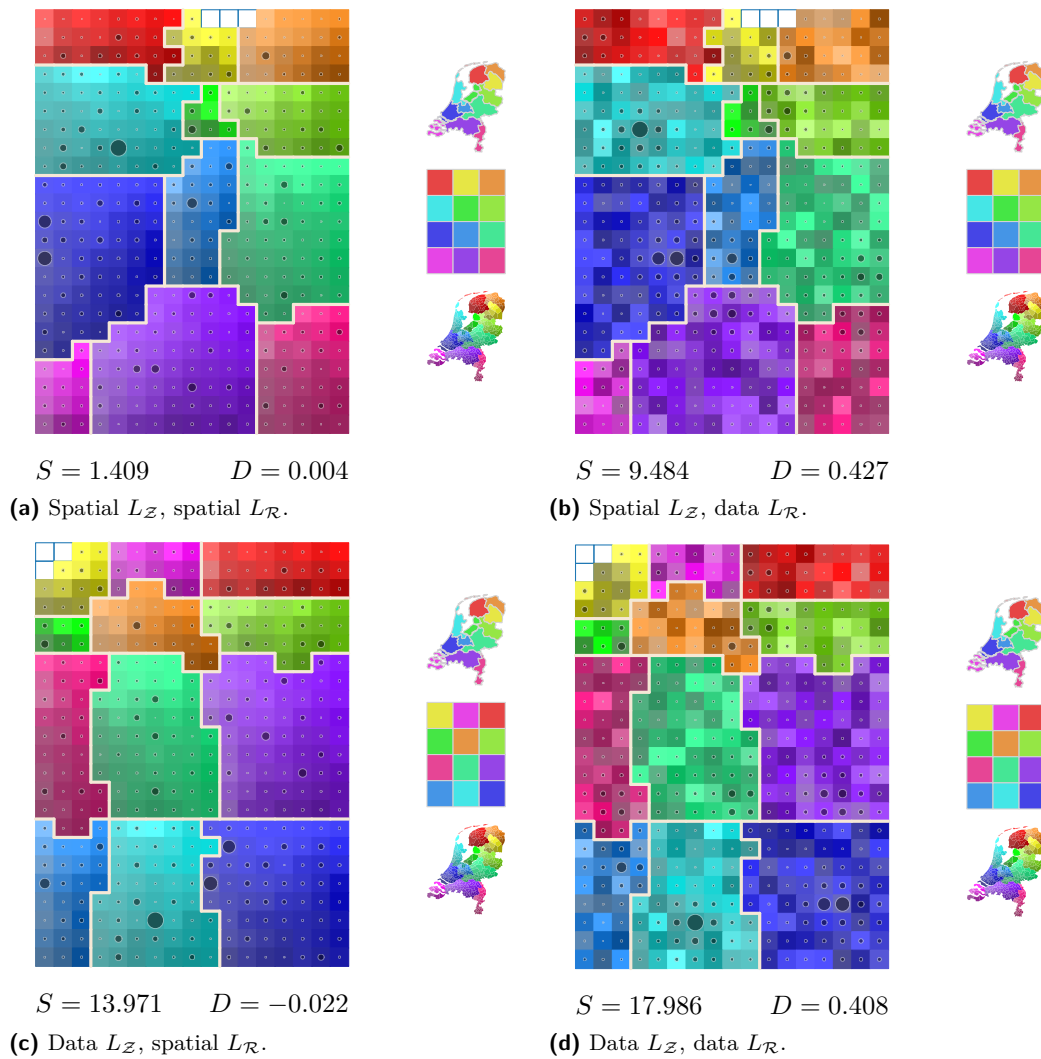


■ **Figure 11** Hierarchical grid map for EW. With each, the smaller maps represent from top to bottom: the zones Z , the zone layout L_Z , and the regions R .

To compute the hybrid layout, we first compute a data or spatial layout L_Z on a coarse grid for the zones. The next step is then to compute the final, region-level layout L_R .

The difficulty lies with zones representing different numbers of regions. We want to obtain an outline for each zone in which to place and rearrange its regions. To this end, we leverage the algorithm for spatial layouts, by creating an artificial “map” based on the zone-level layout. Refer to Figure 10 for illustrations. We scale the regions of each zone z_i such that it exactly fits the tile, translating it such that it is centered within the tile. Subsequently, we scale the regions further by a factor $\lambda < 1$. These steps give us the artificial map, for which we compute the spatial layout L_s of all regions. If the region-level was assigned to be a spatial layout, we are done: L_s is the final layout L_R . If this level was assigned to be a data layout, then we compute a data layout for each zone z_i separately, using the selected tiles for z_i in L_s . The combination of these data layouts per zone then yields the final layout L_R .

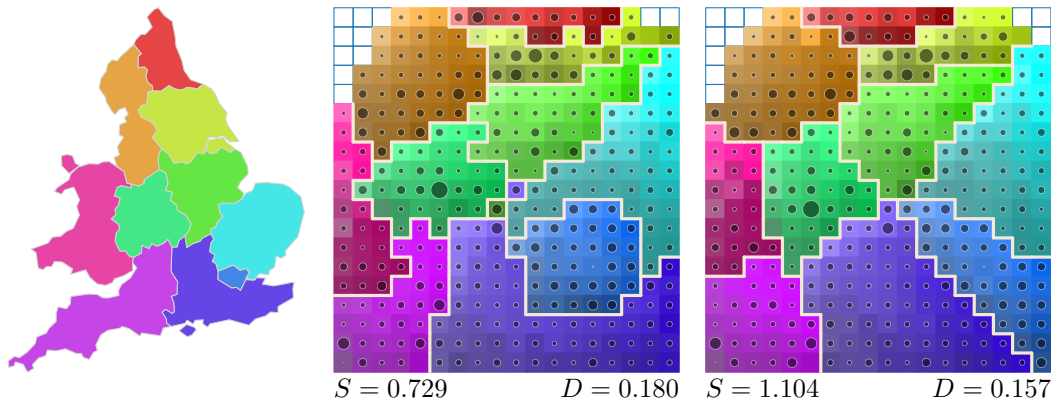
The second scaling step with λ is done to ensure that each zone is represented by a compact, connected set of tiles. The parameter mostly needs to be sufficiently small, we use $\lambda = \frac{1}{W+H}$. For λ to be closer to 1, the resulting zones could potentially preserve more of the shape of the original geography, at the expense of disconnected zones. However, due to the limited space available in the grid in computing the subsequent region-level layout, the gain is limited; when a data layout is used on the region level, the semantics of the preserved shape are lost and may even be misleading.



■ **Figure 12** Hierarchical grid maps for NL. With each, the smaller maps represent from top to bottom: the zones \mathcal{Z} , the zone layout $L_{\mathcal{Z}}$, and the regions \mathcal{R} .

Results. We demonstrate our techniques on our two hierarchical datasets in Figures 11 and 12. We assign each zone a hue, using a saturation and brightness gradient for that hue to color the regions within the zone. We render the boundary between tiles of different zones with a thicker line, to make it easier to identify zones. In both cases we see that our measures indicate that the spatial-spatial layout (a) has the lowest spatial distortion S , while having low data correlation D . The data-data layout (d) achieves the exact opposite.

Interestingly, for the EW dataset, the spatial-data layout (b) has worse spatial correlation than the data-spatial layout (c). Our measures, based on adjacent tiles, are sensitive to local structures being removed in reshuffling each zone, and measure only some form of spatial distortion for the relatively few tiles along the zone boundaries. The better higher-level spatial structure of (b) is not captured as much by our spatial distortion measure S , though it does seem to capture this structure for the NL dataset. Furthermore, it may have quite an influence on what an analyst may understand from the map. That is, it may be easier to still identify and relate the zones in a spatial-data layout, compared to doing so for regions within a zone, that are necessarily harder to identify due to their sheer number.



■ **Figure 13** Center: the spatial layout for the EW data, based only on the regions \mathcal{R} . Note the irregular zones and two disconnected zones. Right: a hierarchical layout computed by scaling regions towards their zone's geographical center.

For region-level data layouts $L_{\mathcal{R}}$ (b) and (d) we can observe uneven coloring in zones of rather similar data values, indicating unnecessary spatial distortion. Hence, instead of using a pure data layout, we could use a hybrid data-to-spatial layout with little slack. As we established in the previous section, such a layout will essentially retain the same data correlation and improve the spatial correlation significantly.

Alternative for spatial zone-level layouts. When using a spatial layout on the zone level, we aim to preserve the geographic relations of the zones. As such, the transformed map we compute in fact aims to resemble to original geography. We could indeed use the original geography – effectively using the non-hierarchical spatial layout for the entire map. Yet, this does have a different effect, as zones are now not represented as compactly. As an example, compare Figure 11(a) to Figure 13 (center): we observe the irregularly shaped zones in the latter, and even two zones that are not represented contiguously.

To promote compactness of zones, we could also opt to shrink the regions of a zone towards the zone's original, geographic centroid. Such an approach avoids the need to compute a zone-level layout. It naturally produces outlines roughly forming a Voronoi diagram of the centroids of the zones; see Figure 13 (right). However, it breaks the grid-like layout in our proposed solution which may help identifiability.

6 Discussion

We presented two approaches which allow us to combine spatial and data aspects in a grid map. Our hybrid grid maps use simulated annealing to effectively improve spatial or data correlation for a layout that is primarily based on the respective other aspect. Our experiments show that already a low amount of slack improves the other measure significantly, while not lowering the quality of the layout much with respect to the original measure – effectively also indicating that Tobler's law is indeed applicable here. Layouts with a large slack are usually inferior to layouts with a small slack computed from the other extreme. We believe that hybrid grid maps with low slack are not only visually pleasing, but also highlight patterns clearly and as such demonstrate the potential of our technique. It is, however, challenging to determine the cause of the complex patterns that arise in the grid maps: the patterns might be complex due to the complex nature of the problem, or due to imperfect grid allocation. Investigating this may be an interesting avenue for future research.

Furthermore, we introduced a controlled way of combining spatial and data aspects, by leveraging a hierarchy, assigning each level of the hierarchy to use a data layout or a spatial layout. This more enforced structure of mixing spatial and data aspects may aid in interpreting the resulting hybrid grid maps, as the zones act as logical units that stay together for higher-level patterns. Although we heuristically achieve a notion of connectivity by choosing a sufficiently small λ , we do not enforce this constraint. In future work, extra constraints may be added in the spatial layout algorithm to ensure connectivity, along the lines of the work by Validi et al. [23]. Though hierarchical grid maps readily generalize to hierarchies of multiple levels, we expect that repeatedly changing from data to spatial layouts and vice versa may result in layouts that are too complex to be understood.

Measures. The quality measures that we used to assess spatial and data quality encapsulate our main goal of creating smooth color and data changes along the grid.

It would be interesting for future research to investigate how well these notions of spatial and data correlation match the expectations of a user, or how effective different measures are for predicting task performance on hybrid grid maps and hierarchical hybrid grid maps.

Usability. The main question looking forward is whether such hybrid layouts indeed help synthesizing a mental model of the data. That is, can a human analyst effectively work with hybrid layouts? With these methods we have shown that it is now possible to create such hybrid layouts, and hence these questions may now be further researched.

Intuitively, it seems that spatial layouts with some improved data correlation are useful. It erases some of the “noise” in the data dimension at the cost of slight distortion of space – but as grid maps are inherently distorted, such seems tolerable and inherent in the approach to begin with. The other end, data layouts with some improved spatial correlation, can also be useful and it is feasible to significantly reduce the spatial distortion while still maintaining a layout with high data correlation. Hierarchical hybrid layouts are more constrained, but for that reason also offer more control in creating such layouts with a strong mix of both aspects. They may hence be easier to interpret than basic hybrid layouts with medium levels of slack.

We color the tiles by their spatial location, to communicate spatial correlation. However, we observe that there may be visual bias for regions with large data values or bright colors. When such regions are contrasting their adjacent tiles, this is more apparent and feels more out of place than when less bright or smaller data values are out of place. The visual assessment is further deceived by an element of the coloring scheme we use for non-hierarchical cases. We rotate the hue around a point in the map. Near this point, regions might be close together while having a different hue, and hence seem far apart in the coloring. We aimed to reduce this effect by reducing brightness and saturation near the center. As an alternative to color, interaction could help in identifying regions.

An eventual hybrid grid map may hence require further attention as to how spatial distortion is communicated and how data values are rendered. A prominent question is to what level of detail spatial distortion should be indicated. Standard grid maps often do not communicate their distortion at all. Yet, in a hybrid layout, being able to separate spatial from data patterns and effects may increase the need for indicators of distortions.

References

- 1 Bonnie Berkowitz and Lazaro Gamio. What you need to know about the measles outbreak, February 2015. Accessed January 2023. URL: <https://www.washingtonpost.com/graphics/health/how-fast-does-measles-spread/>.
- 2 Paul Blickle and Sascha Venohr. Dürfen wir vorstellen: Deutschlands Muslime, January 2015. Accessed January 2023. URL: <http://www.zeit.de/gesellschaft/2015-01/islam-muslime-in-deutschland>.

- 3 Mark Bruls, Kees Huizing, and Jarke J. van Wijk. Squarified treemaps. In *Proceedings of the Joint EUROGRAPHICS and IEEE TCVG Symposium on Visualization*, pages 33–42, 2000. doi:10.1007/978-3-7091-6783-0_4.
- 4 Kevin Buchin, Bettina Speckmann, and Sander Verdonschot. Evolution strategies for optimizing rectangular cartograms. In *Proceedings of the 7th International Conference on Geographic Information Science*, LNCS 7478, pages 29–42, 2012. doi:10.1007/978-3-642-33024-7_3.
- 5 Rafael G. Cano, Kevin Buchin, Thom Castermans, Astrid Pieterse, Willem Sonke, and Bettina Speckmann. Mosaic drawings and cartograms. *Computer Graphics Forum*, 34(3):361–370, 2015. doi:10.1111/cgf.12648.
- 6 Ben Casselman and Allison McCann. Where your state gets its money, April 2015. Accessed January 2023. URL: <http://fivethirtyeight.com/features/where-your-state-gets-its-money/>.
- 7 Mark de Berg, Elena Mumford, and Bettina Speckmann. Optimal BSPs and rectilinear cartograms. *International Journal of Computational Geometry & Applications*, 20(2):203–222, 2010. doi:10.1142/S0218195910003268.
- 8 Danny DeBelius. Let’s tessellate: Hexagons for tile grid maps, May 2015. Accessed January 2023. URL: <http://blog.apps.npr.org/2015/05/11/hex-tile-maps.html>.
- 9 David Eppstein, Marc van Kreveld, Bettina Speckmann, and Frank Staals. Improved grid map layout by point set matching. *International Journal of Computational Geometry & Applications*, 25(02):101–122, 2015. doi:10.1142/S0218195915500077.
- 10 Diansheng Guo, Jin Chen, Alan M. MacEachren, and Ke Liao. A visualization system for space-time and multivariate patterns (vis-stamp). *IEEE Transactions on Visualization and Computer Graphics*, 12(6):1461–1474, 2006. doi:10.1109/TVCG.2006.84.
- 11 Wouter Meulemans, Jason Dykes, Aidan Slingsby, Cagatay Turkay, and Jo Wood. Small multiples with gaps. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):381–390, 2016. doi:10.1109/TVCG.2016.2598542.
- 12 Wouter Meulemans, Max Sondag, and Bettina Speckmann. A simple pipeline for coherent grid maps. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1236–1246, 2020. doi:10.1109/TVCG.2020.3028953.
- 13 Patrick A.P. Moran. Notes on continuous stochastic phenomena. *Biometrika*, 37(1/2):17–23, 1950. doi:10.2307/2332142.
- 14 New York Times. How the rulings affect gay couples, June 2013. Accessed January 2023. URL: <http://www.nytimes.com/interactive/2013/06/26/us/scotus-gay-marriage.html>.
- 15 Kenton Powell, Rich Harris, and Feilding Cage. How voter-friendly is your state?, October 2014. Accessed January 2023. URL: <http://www.theguardian.com/us-news/ng-interactive/2014/oct/22/-sp-voting-rights-identification-how-friendly-is-your-state>.
- 16 Ben Shneiderman and Martin Wattenberg. Ordered treemap layouts. In *Proceedings of the IEEE Symposium on Information Visualization*, pages 73–78, 2001. doi:10.1109/INFVIS.2001.963283.
- 17 Aidan Slingsby. Tilemaps for summarising multivariate geographical variation. In *Proceedings of the Workshop on Visual Summarization and Report Generation*, 2018.
- 18 Aidan Slingsby, Mary Kelly, and Jason Dykes. OD maps for showing changes in Irish female migration between 1851 and 1911. *Environment and Planning A*, 46(12):2795–2797, 2014. doi:10.1068/a140112g.
- 19 Aidan Slingsby, Jo Wood, and Jason Dykes. Treemap cartography for showing spatial and temporal traffic patterns. *Journal of Maps*, 6(1):135–146, 2010. doi:10.4113/jom.2010.1071.
- 20 Waldo R. Tobler. A computer movie simulating urban growth in the Detroit region. *Economic geography*, 46(sup1):234–240, 1970. doi:10.1111/j.1467-8306.2004.09402005.x.
- 21 Alex Tribou and Keith Collins. This is how fast America changes its mind, June 2015. Accessed January 2023. URL: <http://www.bloomberg.com/graphics/2015-pace-of-social-change/>.
- 22 Edward R. Tufte. *The Visual Display of Quantitative Information*, volume 2. Graphics Press Cheshire, CT, 1983. doi:10.1075/idj.4.3.12cos.

- 23 Hamidreza Validi, Austin Buchanan, and Eugene Lykhovyd. Imposing contiguity constraints in political districting models. *Operations Research*, 70(2):867–892, 2022. doi:10.1287/opre.2021.2141.
- 24 Marc J. van Kreveld and Bettina Speckmann. On rectangular cartograms. *Computational Geometry: Theory and Applications*, 37(3):175–187, 2007. doi:10.1016/j.comgeo.2006.06.002.
- 25 Jo Wood, Donia Badawood, Jason Dykes, and Aidan Slingsby. BallotMaps: Detecting name bias in alphabetically ordered ballot papers. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2384–2391, 2011. doi:10.1109/TVCG.2011.174.
- 26 Jo Wood and Jason Dykes. Spatially ordered treemaps. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1348–1355, 2008. doi:10.1109/TVCG.2008.165.
- 27 Jo Wood, Jason Dykes, and Aidan Slingsby. Visualisation of origins, destinations and flows with OD maps. *The Cartographic Journal*, 47(2):117–129, 2010. doi:10.1179/000870410X12658023467367.
- 28 Jo Wood, Aidan Slingsby, and Jason Dykes. Visualizing the dynamics of London’s bicycle-hire scheme. *Cartographica: The International Journal for Geographic Information and Geovisualization*, 46(4):239–251, 2011. doi:10.3138/carto.46.4.239.

Benchmarking Regression Models Under Spatial Heterogeneity

Nina Wiedemann¹  

Institute of Cartography and Geoinformation, ETH Zürich, Switzerland

Henry Martin  

Institute of Cartography and Geoinformation, ETH Zürich, Switzerland

René Westerholt  

Department of Spatial Planning, TU Dortmund University, Germany

Abstract

Machine learning methods have recently found much application on spatial data, for example in weather forecasting, traffic prediction, and soil analysis. At the same time, methods from spatial statistics were developed over the past decades to explicitly account for spatial structuring in analytical and inference tasks. In the light of this duality of having both types of methods available, we explore the following question: Under what circumstances are local, spatially-explicit models preferable over machine learning models that do not incorporate spatial structure explicitly in their specification? Local models are typically used to capture spatial non-stationarity. Thus, we study the effect of strength and type of spatial heterogeneity, which may originate from non-stationarity of a process itself or from heterogeneous noise, on the performance of different linear and non-linear, local and global machine learning and regression models. The results suggest that it is necessary to assess the performance of linear local models on an independent hold-out dataset, since models may overfit under certain conditions. We further show that local models are advantageous in settings with small sample size and high degrees of spatial heterogeneity. Our findings allow deriving model selection criteria, which are validated in benchmarking experiments on five well-known spatial datasets.

2012 ACM Subject Classification Computing methodologies → Concurrent algorithms

Keywords and phrases spatial machine learning, spatial non-stationarity, Geographically Weighted Regression, local models, geostatistics

Digital Object Identifier 10.4230/LIPICs.GIScience.2023.11

Supplementary Material *Software (Source Code)*: https://github.com/mie-lab/spatial_rf_python; archived at `swb:1:dir:f7fc9b237a4e80f882b1785718ab884c2770ac66`

Acknowledgements We would like to thank Martin Raubal for the fruitful discussions about the project.

1 Introduction

The success of machine learning and artificial intelligence in recent years has sparked considerable interest in respective methods also in GIScience, and has led to a general proliferation of spatial data science [24]. While spatial statistics used to carefully address the special nature of spatial data, spatial data science often involves the direct application of (global) machine learning models to spatial data without explicitly modeling spatial properties. Nevertheless, these models oftentimes provide successful inferences on test data. Yet, spatial data may be subject to complex confounders including spatial heterogeneity, which is the focus of this paper. Currently, there is no comprehensive review available

¹ Corresponding author



that would show when global non-linear machine learning models can or should be used for interpolation or prediction tasks on spatially heterogeneous data without producing misleading or wrong results. While there exist analyses on the effectiveness of methods to deal with spatial autocorrelation [2], no such benchmarking has been done regarding spatial heterogeneity, either originating from non-stationarity of the actual process or from spatially heterogeneous exogenous noise. In this work, we benchmark the performance of global (machine learning) and local spatial regression models for the prediction of unseen test data that is subject to various kinds of heterogeneity. We simulate spatial heterogeneity with synthetic data in order to derive recommendations about the suitability of model types for specific heterogeneity-related scenarios. We finally validate our model selection criteria through experiments on several real-world datasets. The following two sub-sections briefly outline the state of the art as well as our contribution in more detail, before we present the experiments and our results.

1.1 Related work

Statistical learning methods have been adapted to geospatial data since a long time. A major step towards accounting for spatial heterogeneity has been the proposal of local models, such as Geographically Weighted Regression (GWR) [3, 9]. Next to variants of GWR [17], the idea also inspired adaptations of machine learning models, with spatial versions of Random Forests (RFs) [11, 28] or even Geographically Weighted Artificial Neural Networks [13, 7]. The proposed modifications of machine learning models such as Random Forests include 1) providing spatial coordinates as input [18], 2) deriving spatial features such as the distance from points of interest from the coordinates [14] in order to improve spatial generalization [5], 3) including the observations at nearby samples as covariates [28], and 4) fitting RFs on local subsets of data [11]. While these approaches have been shown advantageous in some situations, a recent study Zhou et al. [31] compared GWR with geographical RFs on health data and actually found that GWR provided better predictions than the more complex RF models, though the generalizability of the results is limited due to the very specific application context.

A common limitation of existing approaches is that the developed methods are usually evaluated on a single or few real dataset(s). The results may therefore be subject to unknown data properties. Synthetic data, in contrast, allows to benchmark methods in a controlled setting. While this solution is implemented, for example, by Beale et al. [2] and Santibanez et al. [26] for the purpose of assessing the effect of varying degrees of spatial autocorrelation, there is a lack of benchmarking with simulated spatial heterogeneity. Fotheringham et al. [10] and Hagenauer et al. [13] validate their methods on synthetic data that were designed to be non-stationary in space, and Finley et al. [8] compare GWR and SVC on non-stationary synthetic data, but they do not systematically vary the non-stationarity. The latter is our point of departure for the following sections.

1.2 Contribution

We evaluate the ability of different models to deal with varying degrees of spatial heterogeneity. Inspired by the work conducted by Comber et al. [4] presenting a route map when to use GWR and two of its variants, we derive model selection criteria from our results on synthetic data. We extend previous findings in three ways: first, in addition to GWR and other linear methods, we also consider Random Forests as non-linear models and compare their performance on non-linear tasks; second, we consider *predictive* performance instead of

analysis in order to account for overfitting behavior; third, and most importantly, we provide a detailed analysis of model adequacy with respect to spatial non-stationarity and signal-to-noise ratio. To achieve this, we propose a synthetic data-generating process that allows to systematically vary the degree of spatial heterogeneity due to 1) the non-stationarity of the process, and 2) noise. We utilize this framework to compare seven models that are selected to reflect standard approaches that were, to varying degrees, developed to deal with spatial data. By analyzing the model performances in this controlled synthetic setting, we derive recommendations what model is appropriate dependent on the sample density, the spatial heterogeneity and the problem complexity. We validate our model selection criteria by benchmarking the models also on five real geospatial datasets.

2 Methods

We simulate a spatial regression problem with synthetically generated data that are subject to spatial heterogeneity. Spatial heterogeneity in our analysis stems from two effects; on the one hand, the dependence of the dependent variable² Y on the independent variables X may be non-stationary, i.e., the same input may lead to different outputs in different spatial regions. In previous work [10, 13], this was modeled by varying the coefficients β dependent on the coordinates (u, v) ; for example, Fotheringham et al. [10] set $\beta_1 = 1 + \frac{(u+v)}{12}$ and Hagenauer et al. [13] add coefficients with oscillating spatial distribution based on trigonometric functions. On the other hand, spatial heterogeneity may be caused by differences in the variance of the errors (and thus by noise). To understand the effect of the signal-to-noise ratio in spatial data subject to spatial heterogeneity of both types, we propose to vary the noise and the level of non-stationarity over space and to compare models on both linear and non-linear problems on test data.

2.1 Data-generating processes (DGPs)

One of our investigated DGPs represents a linear relationship of Y on k independent variables $x_j (j \in [1..k])$. It is given as

$$y_i = \sum_j^k \beta_j(u_i, v_i) \cdot x_{ij} + \epsilon(u_i, v_i), \quad (1)$$

where x_{ij} is the j -th feature of the i -th sample, (u_i, v_i) are the coordinates of the i -th sample, and $\beta_j(u_i, v_i)$ is the location-dependent coefficient. $\epsilon(u_i, v_i)$ is the noise that may also be heterogeneous across space. The definition of β and ϵ will be given in detail in Section 2.1.1 and Section 2.1.2 respectively.

We also implement a non-linear DGP in order to analyze the model performances under the regime of a more complex phenomenon. The function is constructed such that there are interactions between variables and non-linear effects of single variables, and the terms are weighted with the non-stationary coefficients β :

$$\begin{aligned} \tilde{y}_i = & \beta_1(u_i, v_i) \cdot x_{i1}^2 \cdot \sin(x_{i2}) + \beta_2(u_i, v_i) \cdot \sin(x_{i2}) \cdot x_{i4} \\ & + \beta_3(u_i, v_i) \cdot x_{i5} \cdot \log(x_{i3}^2) + \beta_4(u_i, v_i) \cdot x_{i4}^2 \cdot \cos(x_{i2}) \\ & + \beta_5(u_i, v_i) \cdot x_{i1}^2 \cdot x_{i4} \cdot x_{i5} + \epsilon(u_i, v_i). \end{aligned} \quad (2)$$

² Throughout this paper, we use capital characters for vectors and matrices and non-capitalized characters for referring to scalar terms.

In both scenarios, we construct n samples with pairs of geographic coordinates (u_i, v_i) and k attribute values x_{ij} . The coordinates are drawn from a uniform distribution $\mathcal{U}(-1, 1)$. In contrast to related work, we did not use coordinates on a regular grid in order to better mimic a realistic situation with irregular local clustering and dispersion patterns of observation sites. The independent variables X are assumed to be subject to spatial autocorrelation since we aim to simulate realistic spatial data. This is modeled by left-multiplying a vector of uniform random data X' by the so-called spatial autoregressive (SAR) generating operator³ [16], that is, as $X = (I - \rho W)^{-1} X'$, where W is the weight matrix, computed as the inverse distances of the 20 nearest neighbors. After observing that the average spatial autocorrelation, measured using Moran's I , is around 0.3 in the considered real datasets, we calibrate the autoregressive parameter ρ such that the resulting values yield Moran's I values of around 0.3 accordingly ($\rho = 0.75$).

2.1.1 Non-stationary coefficients β

In contrast to previous work assuming a complete variation of the coefficients [10, 13], we argue that with many types of real-world processes, it would be more reasonable for the coefficients to vary around a constant value c_j . To simulate this, we frame spatial non-stationarity as an additive factor to the underlying coefficient c_j , and quantify its strength with a factor λ . The coefficients used are thus composed of the constant coefficient c_j and the spatial variation $\hat{\beta}_j(u_i, v_i)$:

$$\beta_j(u_i, v_i) = c_j + \lambda \cdot \hat{\beta}_j(u_i, v_i).$$

The spatial variation $\hat{\beta}$, in turn, is modeled based on trigonometric functions and thus in a similar fashion as presented in [10, 13]:

$$\hat{\beta}_j(u_i, v_i) = \sin(u_i \cdot 2\pi + j) + \cos(v_i \cdot 2\pi + j).$$

Since the coordinates are drawn from $\mathcal{U}(-1, 1)$, this definition of $\hat{\beta}$ leads to two cycles of the sine and cosine functions in x and y direction. Furthermore, the spatial variation is shifted by j for the j -th coefficient to ensure that the spatial heterogeneities attached to the coefficients are not all the same. The final coefficients $\beta_j(u_i, v_i)$ with weak ($\lambda = 0.2$) and strong ($\lambda = 0.5$) non-stationarity are shown in Figure 1.

2.1.2 Spatial heterogeneity of the errors ϵ

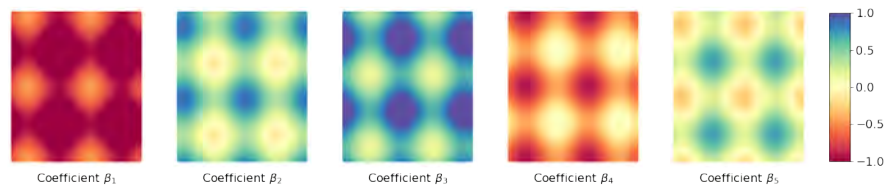
Not only β -coefficients but also the error terms can vary across space. A heterogeneous spatial distribution of the noise ϵ increases the difficulty of distinguishing signal from noise. The spatial distribution may thereby either be similar to one of the coefficients (i.e., also trigonometric) or different. Let σ be the average noise strength similar to the non-stationarity effect size λ as defined in Section 2.1.1. Using this, we consider three scenarios for varying the error terms:

$$\epsilon \sim \mathcal{N}(0, \sigma), \tag{3a}$$

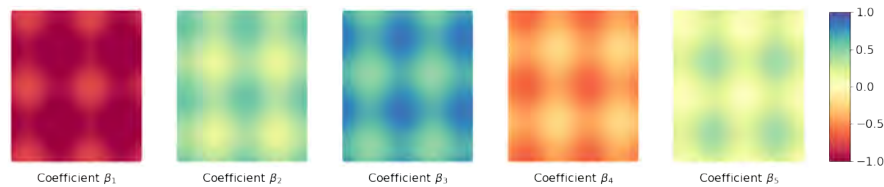
$$\epsilon(u_i, v_i) \sim \mathcal{N}(0, \hat{\sigma}(u_i, v_i)) \quad \text{with} \quad \hat{\sigma}(u_i, v_i) = \sigma \cdot (\sin(u_i \cdot 2\pi) + \cos(v_i \cdot 2\pi) + 1), \tag{3b}$$

$$\epsilon(u_i, v_i) \sim \mathcal{N}(0, \hat{\sigma}(u_i, v_i)) \quad \text{with} \quad \hat{\sigma}(u_i, v_i) = \sigma \cdot (0.5 \cdot (u_i + v_i) + 1). \tag{3c}$$

³ See also <https://r-spatial.github.io/spatialreg/reference/invIrM.html>

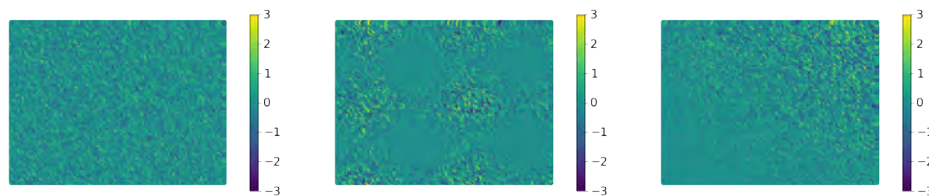


(a) Coefficients with strong spatial heterogeneity ($\lambda = 0.5$).



(b) Coefficients with weak spatial heterogeneity ($\lambda = 0.2$).

■ **Figure 1** Spatial non-stationarity is simulated as a trigonometric spatial variation of the coefficients β . The factor λ determines the overall strength of the non-stationarity.



(a) Uniformly distributed noise.

(b) Heterogeneous (trigonometric).

(c) Heterogeneous (linear).

■ **Figure 2** Varying the spatial distribution of the variance of the errors ϵ . We simulate three scenarios: uniformly distributed noise ϵ , one that follows a similar distribution as the non-stationary process (i.e., trigonometric), and one that follows a different distribution (linear).

Equation 3a refers to a scenario with uniformly distributed noise. This scenario does not incorporate spatially varying errors. Equation 3b describes error terms that are heterogeneous in the sense that their variance oscillates trigonometrically around σ , depending on their spatial locations. The last scenario presented in equation 3c is also spatially varying but based on a diagonal linear trend over the map. Respective noise maps created under the scenarios outlined are illustrated in Figure 2.

2.2 Regression models

We consider linear and non-linear, global and local models suitable for regression tasks. Figure 3 provides an overview of their properties. In the following, let $X \in \mathbb{R}^{n \times m}$ denote the m -dimensional feature matrix of n samples, and let $Y \in \mathbb{R}^n$ be the dependent variable that is to be predicted from X .

2.2.1 Ordinary Least Squares and a global spatial model

We employ two linear global types of regression models. One of these is the Ordinary Least Squares (OLS) model, which assumes a linear dependency of Y on X . It is given as

$$Y = X\beta + \epsilon,$$

11:6 Benchmarking Regression Models Under Spatial Heterogeneity

Ability to deal with...	OLS	SAR	GWR	RF	RF (coordinates)	Spatial RF	Regression Kriging
Non-linear data	X	X	X	✓	✓	✓	✓
Spatial autocorrelation	X	✓	✓	X	X	X	✓
Non-stationarity	X	X	✓	X	(✓)	✓	✓

■ **Figure 3** Overview of the compared models' abilities to handle non-linearity, their consideration of spatial autocorrelation, and their respective suitability for non-stationarity.

with ϵ being the error term and $\beta \in \mathbb{R}^m$ denoting the coefficients. In OLS, the coefficients can be estimated using matrix inverse and multiplication: $\beta = (X^T X)^{-1} X^T Y$. The intercept can be included in this model through a column vector of ones added to the feature matrix, which yields $X \in \mathbb{R}^{n \times m+1}$ and $\beta \in \mathbb{R}^{m+1}$. Note that applying OLS on spatial data is not generally advisable since it assumes that the samples are independent. This is not the case with (geo)spatial data because these are often taken from shared contexts, originate from processes with endogenous spatial dispersal mechanisms, or may be driven by spatially structured covariates. We nevertheless include OLS in our comparison as it is widely used as a yardstick against which to assess the usefulness of spatially explicit methods.

The second global linear method tested is the Spatial Lag in X model (SLX). This model takes into account spatially lagged independent variables and is given as

$$Y = \rho W X + X \beta + \epsilon,$$

where W is the spatial weights matrix that is computed as the inverse distance of the 20 nearest neighbors (see DGP), and ρ is the spatial coefficient. The estimation of β and ρ can be solved by adding the spatially-lagged X as additional covariates, and estimating two sets of coefficients for X and $W X$ respectively.

2.2.2 Geographically Weighted Regression

Although Geographically Weighted Regression (GWR) was proposed for the analysis (not prediction) of spatial data, it is a suitable local model to account for non-stationarity in regression problems. GWR follows the standard linear regression framework but assumes that the coefficients β are dependent on locations (u_i, v_i) . The model specification is given as

$$y_i = \sum_j \beta_j(u_i, v_i) X_{ij} + \epsilon.$$

In GWR, the local coefficients are estimated by building local models around each sample including only the spatial neighbors within a bandwidth. The latter can either be *fixed* (i.e., a pre-set distance) or *adaptive* (i.e., varying in space). The bandwidth is optimized by means of the golden-section search algorithm based on the Corrected Akaike Information Criterion (AICc) or with cross-validation (CV). Here, we tune a *fixed* bandwidth with the AICc criterion and use an exponential kernel. Our analysis aims to benchmark established local and global models on a synthetic (single-scale) task. We, therefore, use the original GWR specification but do not consider variants of the model such as multi-scale GWR [10].

2.2.3 Random Forest Regression models

Random Forests (RFs) are established machine learning models for regression tasks and have been shown to be very successful for a wide range of applications. We choose RFs as the main non-linear model in our experiments since it is arguably most prominent in

spatial applications and does not require extensive parameter tuning. An RF is formed as an ensemble of decision trees that can learn arbitrary non-linear relations. The prediction of an RF is the average over the tree-wise outputs. We use the implementation provided through the `scikit-learn` [23] package.

To give RFs the ability to learn spatially non-stationary processes, a simple approach is to include the geographic coordinates as covariates [18]. We denote this RF-variant by *RF (coordinates)* in the following. In general, this approach is not recommended, since such a model is not applicable to other spatial regions [5]. However, we only regard regression within the same region here.

2.2.4 Spatial Random Forests

Aside from simply extending non-linear models by adding geographic coordinates or spatial features as covariates, another option is to fit them locally, as a non-linear counterpart to GWR. Similar to [11], we implement this approach for RFs. To provide a local yet efficient approach, we exploit the bootstrapping nature of RFs and fit a fixed number of spatially-disjoint decision trees. The decision trees are rooted in the cluster centers of K-Means clustering applied to the dataset. At test time, the prediction for a test sample is given by the weighted average of tree-wise predictions, where the weights are defined by the inverse distances of the test sample to the root of each tree respectively. While Georganos et al. [11] proposed a weighting of the spatial-RF and the global-RF predictions, we set the weight to 1 for a fair comparison between global and local models. Our version of spatial RFs is made available as an open-source package⁴. We validated that our spatial RF achieves similar performance as the implementation by Georganos et al. [11] and found that it is actually superior in 65.7% of all simulated scenarios and under *ceteris paribus* conditions.

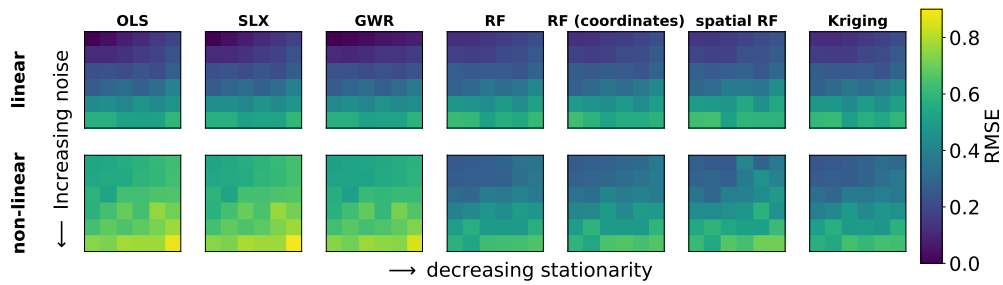
2.2.5 Kriging

Another method that we employ is Kriging. This method is a well-known approach for interpolating geospatial data. A suitable variant for regression tasks is so-called Regression Kriging, which corresponds to universal Kriging with external drift. Regression Kriging essentially tackles (possibly non-linear) regression problems by fitting an arbitrary (global) regression model on the data and then applying Kriging on the *residuals*. Here, we use an RF as the base regressor in order to achieve maximum comparability to the global RF models, and employ the Kriging implementation offered in the `pykrige` package [20]. All Random Forest-based models are fitted with 100 base estimators and a maximum tree depth of 30. Increasing the number of estimators to 150 did not yield any significant improvements. We did not tune other parameters for a fair comparison. For GWR and spatial RFs, the bandwidth is tuned on validation data.

2.3 Experimental setup

We construct synthetic data following the DGPs described above, and evaluate the seven regression models in each scenario. The data is thereby randomly split and each model is trained on 90% of the data and tested on the remaining 10%. To study the effect of the sample size, we generate four datasets with $n = 100$, $n = 500$, $n = 1000$, and $n = 5000$ samples respectively, and $k = 5$ attributes for each sample. Our DGPs allow to compare model performances subject to varying degrees of non-stationarity (λ) and of the variance of

⁴ https://github.com/mie-lab/spatial_rf_python



■ **Figure 4** Results on the synthetic dataset (1000 samples). Performance in general decreases with noise and with the degree of non-stationarity (lowest performance in the bottom right of each plot). On linear data, GWR can account for non-stationarity, in contrast to other models. A random forest is better suited for non-linear phenomena, but spatial (locally fitted) RFs do not provide any benefits in these scenarios.

error terms (σ). In our experiments, we systematically vary the spatial non-stationarity by setting the factor λ to values between 0 and 0.5 (see Section 2.1.1). Furthermore, we vary the signal-to-noise ratio by setting σ to values between 0 and 0.5, where $\sigma = 0.5$ corresponds to a low signal-to-noise ratio (i.e., strong noise).

3 Results and discussion

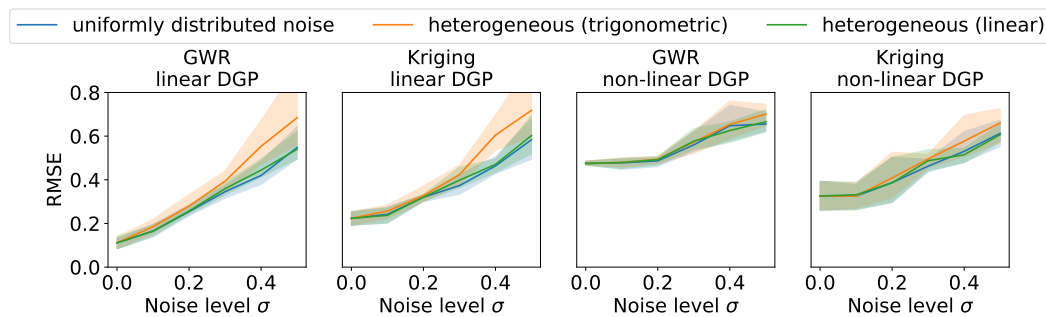
In the following, we first compare the performances of the models on our synthetic dataset, then derive recommendations for model selection, and finally validate these recommendations in experiments on five real-world datasets.

3.1 Results obtained from synthetic data

The model performances in terms of test-data RMSE are visualized in Figure 4, divided by data generating function (row), spatial non-stationarity (x-axis), and noise level (y-axis). Only the scenarios with 1000 samples and uniformly distributed noise ϵ are shown. As expected, the performance generally decreases with higher noise levels or higher degrees of spatial heterogeneity (see highest RMSE in the bottom right corner of each scenario depicted in Figure 4). For the linear DGP, one can clearly see the superiority of GWR in dealing with locally varying spatial data, as it is indeed very robust to the adjusted spatial heterogeneity. The linear models (GWR, OLS, and SAR) are also clearly better at dealing with noise in linear regression tasks, whereas non-linear regressors such as Random Forests may struggle from overfitting. However, the latter picture changes when considering a non-linear function. The non-linear models yet generally struggle more with spatial non-stationarity than their linear counterparts. Surprisingly, spatial RFs are consistently outperformed by other models for the linear case, probably due to overfitting local models on the limited number of samples. The figure further indicates that a spatial RF is also not the best model when it comes to non-linear scenarios, though better than GWR. In this case, the problem may be underfitting, given the lower number of samples that are fed into each local model.

3.1.1 Effect of the spatial heterogeneity of the errors

As explained in Section 2.1.2 we additionally simulate different distributions of the variance of the errors ϵ (see Figure 2). Figure 5 visualizes the RMSE for GWR and Regression Kriging by the noise level. The outcomes obtained for degrees of non-stationarity $\lambda \in \{0.3, 0.4, 0.5\}$



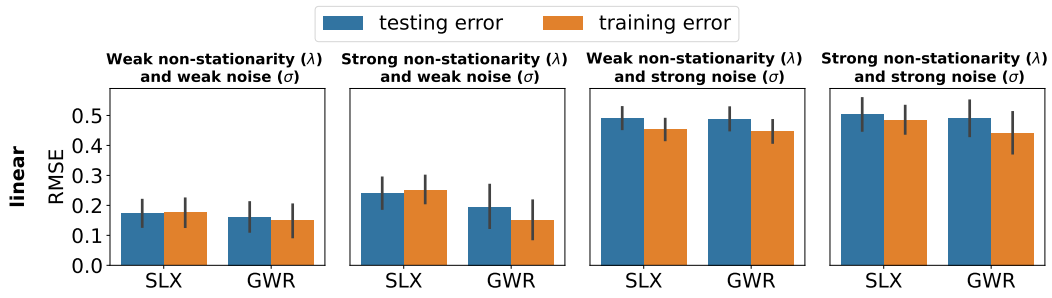
■ **Figure 5** The average RMSEs with their 95% confidence intervals for varying noise levels (500 samples, averaged over scenarios with high degrees of non-stationarity). The RMSE is highest if the noise varies in the same fashion as the coefficients (heterogeneous – trigonometric) for the linear DGP. For the non-linear DGP the noise pattern has no significant influence.

are thereby averaged for obtaining an easier-to-interpret picture, so the blue lines (uniformly distributed noise) in Figure 5 correspond to the right part of the squares in Figure 4. In general, the type of distribution only has a minor effect compared to the average noise level, in particular for the non-linear GDP. However, at stronger noise levels, the scenario with trigonometrically varying noise is clearly the most difficult. Additionally, the variance of the RMSE increases in that case. Since the non-stationarity of the coefficients β is also modeled trigonometrically, these findings indicate that the models particularly struggle to distinguish signal from noise if the variance of the errors is distributed similarly to the non-stationarity.

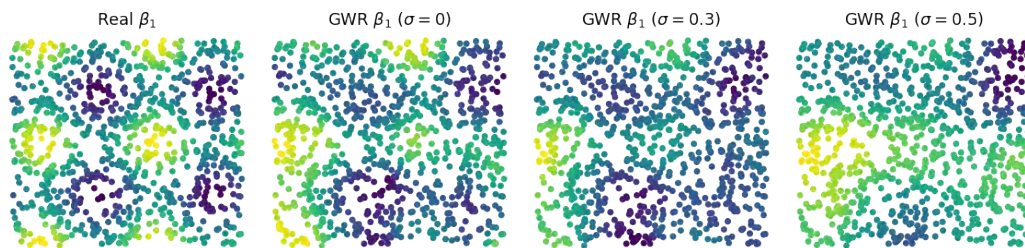
3.1.2 Comparing training and test errors

GWR and related local models are oftentimes only evaluated in terms of their fit to the input data, and not by means of inference on unseen data. Since GWR is based on linear models, the risk of overfitting is considered low, and evaluating the fit on test data is deemed unnecessary. Here, we make the case for evaluating models in terms of their predictive power, since even local linear models may overfit due to their higher number of parameters, and because local sample sizes are often small. To justify this argument empirically, we compare the RMSE on training and test data in our experiment. We find that Random Forest-based models (including Regression Kriging as we base it on RF) generally achieve very small training errors ($\text{RMSE} < 0.01$), which is expected since the individual decision trees overfit on the training data and only the boosting approach leads to good test performance. In Figure 6 we therefore only compare the results for SLX and GWR to showcase the danger of overfitting even linear models when they are local. Here, we consider $\lambda \in \{0, 0.1, 0.2\}$ as “weak non-stationarity” and $\lambda \in \{0.3, 0.4, 0.5\}$ as “strong non-stationarity”, $\sigma \in \{0, 0.1, 0.2\}$ as “weak noise” and $\sigma \in \{0.3, 0.4, 0.5\}$ as “strong noise”. The results are averaged over these scenarios for $n = 1000$ samples. Figure 6 shows that SLX as a global linear model hardly overfits on the data, whereas for GWR, which has considerably more parameters than global linear models, the training and test errors indeed diverge in some scenarios. For example, when there is strong non-stationarity but weak noise, the test error of GWR is 31% higher than its training error. This demonstrates the necessity to validate models on test data when employing them in predictive instead of purely analytical scenarios.

Additionally, overfitting may even lead to misinterpretations of analytical results of GWR, such as the visualization of the estimated coefficients on a map. The effect of overfitting on the spatial interpretation is application-dependent, but we exemplify the problem in



■ **Figure 6** Comparing training and test errors in different scenarios. Even linear models such as GWR show overfitting behaviour, i.e., a lower test than train score, if there is either noise or non-stationarity in the data.

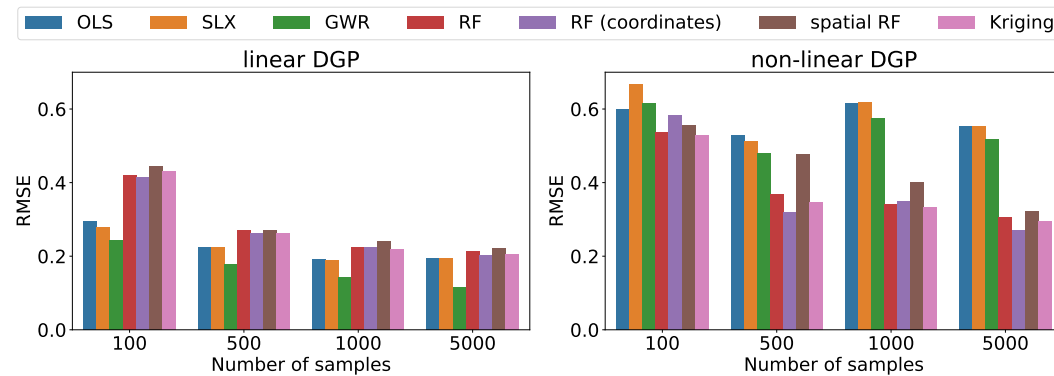


■ **Figure 7** Comparing the GWR-estimated coefficients to the real coefficient for different signal-to-noise ratio. With noisy data, the spatial interpretation can be distorted.

Figure 7. The figure shows the true spatial variation of one coefficient β_1 in synthetic data ($n = 1000$) with moderate spatial heterogeneity ($\lambda = 0.3$), as well as the distribution of its estimate obtained with GWR. With decreasing signal-to-noise ratio, the spatial pattern of the estimated coefficient is perturbed. The pattern for $\sigma = 0.5$ indicates a single area with high β_1 on the left side of the region, in contrast to the true trigonometric pattern. This shows the potential for misinterpreting the results of a model with a bad fit to the data and calls for validation on test data before spatial analysis and interpretation. Of course, there is no unequivocal and generally agreed definition for when a model is overfitting, and overfitting may not be problematic as long as the test performance is sufficiently high. However, the *interpretation* of coefficients should be considered with caution in such case. For example, one could only analyze the coefficients of local models that were fit on a sufficient number of samples.

3.2 Proposed criteria for model selection

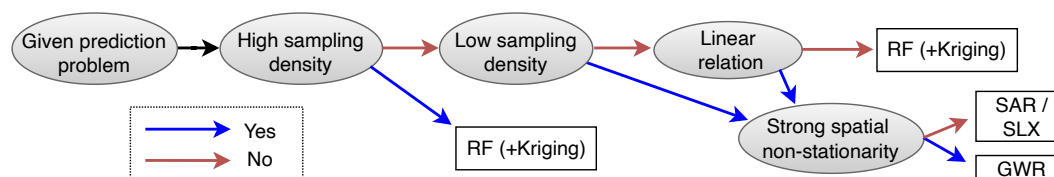
Our experiments on synthetic data allow to derive recommendations for choosing a model, dependent on the prediction task and on data availability. In general, the results in Figure 4 render linear models such as SLX most suitable for the linear DGP, with clear advantages of GWR in non-stationary scenarios. In contrast, Random Forest-based models are superior in the case of a non-linear DGP, while local RFs do not seem to provide many benefits. However, in real-world scenarios, the DGP is usually expected to be neither perfectly linear nor as complex as our non-linear scenario. It is therefore worthwhile to consider further factors such as the sample size. For this purpose, we analyze scenarios with strong non-stationarity ($\lambda \in \{0.3, 0.4, 0.5\}$) and weak noise ($\sigma \in \{0, 0.1, 0.2\}$) by sample size in Figure 8. Note that



■ **Figure 8** Comparing performance by the number of samples. The figure shows the average RMSE over all scenarios with strong non-stationarity ($\lambda \in \{0.3, 0.4, 0.5\}$) and weak noise ($\sigma \in \{0, 0.1, 0.2\}$).

the samples are constructed by *infill sampling* in a fixed spatial region, implying that a higher n leads to a higher sample *density*. For $n = 1000$ samples, the results correspond to the average over the top-right quadrant of each square in Figure 4.

As Figure 8 shows, RF models perform similarly well on linear data in scenarios with high sample density, whereas GWR is almost on-par for non-linear data when only 100 samples are provided. This observation leads us to derive the model selection tree presented in Figure 9: If there is clearly a high sample density over space, RFs should be used, whereas linear models are advisable in scenarios with very low sample density, or if the phenomenon is expected to show linear relations. Since the value of a “high” or “low” sampling density is application-dependent, this criterion must be decided on the basis of an analysis of the local number of samples, e.g., by the number of samples within the set range in a semivariogram. In scenarios with high non-stationarity, Kriging or spatial features in the RF are beneficial. Global RFs should be tested in any case, in order to validate the necessity of local models. It must be noted, however, that our analysis does not consider big data scenarios, where RFs may still perform well but would need to be replaced by more memory-efficient methods such as stochastic gradient descent.



■ **Figure 9** Proposed criteria for model selection. The model choices were derived from experiments with synthetic data of varying non-stationarity, sample size, and DGP (linear vs non-linear).

3.3 Results based on real-world data

We experiment with five benchmark datasets that have been used in previous work on spatial data analysis and prediction, e.g. [19, 22, 14, 1]. The following sub-section first introduces these datasets. Afterward, we discuss the results obtained.

3.3.1 Datasets

There are five real-world, publicly available datasets that we employ for validation:

- The **California housing dataset**⁵ was generated from the 1990 California census. Our goal is to predict the median house price from the location and seven other variables, such as the size and number of rooms, age, income, and population size. The number of bedrooms is missing for 1% of the houses and we omitted those respective records.
- The **Atlantic mortality dataset**⁶ captures county-level mortality rates from 2010–2012, from which we have extracted only one year’s worth of data for our purposes. Rates of smoking and poverty, as well as PM25, SO2, and NO2 levels provided as annual means are utilized as covariates.
- We further use a dataset on **deforestation rates**⁷ that was published by Santos et al. [27]. The dataset provides annual deforestation rates from 2000 to 2010 for 2418 grid cells (single values averaged over 10 years). The deforestation rate is to be predicted from 35 further variables about sociodemographics, spatial features, and economic information. The forestation rate is given as four *quantiles*, which is problematic for the framing as a *regression* problem. The results must therefore be taken with a grain of salt.
- The **Meuse river dataset**⁸ is another standard dataset for experimental spatial analysis [25]. It is a rather small collection of soil measurements including copper, cadmium, and zinc. Usually, this dataset is used to predict zinc concentration from the other soil measurements as well as from further contextual information. For preprocessing, we omit the categorical “landuse” variable and two incomplete samples.
- Finally, a dataset on **plant richness** is included that was used for validating spatial random forests⁹. Plant species richness is given for 227 ecoregions in America, and there are 18 covariates with information on topography, land use, human population, and climate.

3.3.2 Results obtained from real-world data

To validate the model selection tree presented in Figure 9 on real-world data, we first compute an indicator for the degree of non-stationarity. The LOSH statistic [21] offers a way to estimate local heterogeneity in terms of a local, spatially-weighted variance estimator. When applying LOSH with a K-nearest-neighbor (KNN) weights matrix (here 20 neighbors), the global average of all LOSH values indicates the average heterogeneity with respect to the sample density. As shown in Table 1, we find LOSH values around 1.0 in the five real-world datasets, where the California housing and the Meuse datasets show lower local heterogeneity (ϕ LOSH of 0.88 and 0.89), and the plants data is subject to stronger local heterogeneity (ϕ LOSH of 1.06). Table 1 further gives the number of samples and the number of covariates k as an indicator of the problem complexity. We then quantified the model performances in terms of RMSE, mean absolute error (MAE), and the R-squared score; however, all metrics yield the same ranking of methods, and we therefore only report the RMSE in Table 1.

⁵ We use the public dataset available from Kaggle: <https://www.kaggle.com/datasets/camnugent/california-housing-prices?resource=download>.

⁶ The data is available from <https://zia207.github.io/geospatial-r-github.io/geographically-weighted-random-forest.html>.

⁷ Data downloaded from <https://github.com/FSantosCodes/GWRFC/tree/master/data>

⁸ The data is included in the R package `sp`: <https://rsbivand.github.io/sp/reference/meuse.html>

⁹ The data is available from <https://blasbenito.github.io/spatialRF/#data-requirements>.

■ **Table 1** Model benchmarking on real-world data. We find that GWR performs better on the Atlantic dataset and the Meuse data, whereas non-linear models yield lower RMSE on datasets with higher sample density as expected (e.g. California housing).

Dataset	Samples	k	\varnothing LOSH	RMSE						
				OLS	SLX	GWR	RF	RF (coord.)	spatial RF	Kriging
Atlantic	666	7	1.00	8.65	8.64	7.14	7.54	7.34	8.18	7.43
California housing	20433	8	0.88	72244	63532	56156	61234	48209	67493	55173
Deforestation	2418	36	1.00	0.83	0.82	0.80	0.66	0.67	0.71	0.66
Meuse	153	11	0.89	51.65	54.80	48.40	68.13	68.13	88.70	64.16
Plants	227	18	1.06	2349	2334	2226	2216	2288	2507	2120

For validating the results, we compare to previous results reported for these datasets, and our scores improve over the ones reported in the data-accompanying tutorials^{6,9}, or achieve comparable results as related work [19].

We confirm previous results that spatial models achieve good results on these spatial datasets. However, RF-based methods perform better on several datasets, in particular, when the sample density is sufficiently high (e.g., California housing) or when the process analyzed is more complex (e.g., predicting quantiles in the deforestation dataset from 36 variables; or predicting plant richness from 18 covariates). A surprising result is the superiority of GWR above other model specifications for the Atlantic dataset (mortality rates) despite the intermediate LOSH value and sample size. This may be due to a rather linear dependency of Y on X , and is in line with previous findings [31]. Our results on real-world data, therefore, show the general applicability of our model selection criteria, but call for further efforts on quantifying spatial non-stationarity and problem complexity in spatial data.

4 Conclusions

While many promising regression methods were developed specifically for spatial data, there is a lack of analysis about the properties of data that render such models superior. We contribute to a better understanding of these conditions with an analysis systematically exploring the effects of non-stationarity, the signal-to-noise ratio, noise heterogeneity, the nature of the DGP (linear/non-linear), and sample size. Based on the experiments, we recommend using (local) linear models such as GWR for addressing problems encompassing a small sample size or strong non-stationarity. Further, we recommend using non-linear models such as Random Forests for prediction tasks involving larger spatial datasets, whereby locations should be fed into the model through additional spatial input features. RFs can further be combined with Kriging to better account for non-stationarity. While the type of data may give some indication of the non-stationarity and complexity, further work is necessary to assess spatial stationarity a priori. Promising avenues may be, for example, exploring spatial stationarity measures as proposed for time series [6], through better understanding localized (and varying) heterogeneity [30] or, alternatively, by controlling for complex forms of stationarity using Moran eigenvector filtering and its variants [29, 12]. We further argue that our results call for an increased significance of prediction for validating model performance. Even if a model is only used for analysis, the validity of the inferred coefficients should be evaluated via test data, since even linear local models are prone to overfitting in spatially structured noisy or non-stationary settings. At the same time, other factors that are not discussed in this work, such as model interpretability, may be important when it comes to model selection and may give preference to linear modeling even though non-linear models may be superior in terms of prediction. Future work could aim to combine the best of both worlds by improving the spatial interpretability of global models such as RFs.

Finally, our analysis is limited in scope regarding the considered properties and models. Follow-up work could put more focus on spatial autocorrelation and its interplay with non-stationarity, or explore other types of non-stationary non-linear relations. Another interesting path that some researchers have started venturing on is to integrate better modern machine learning models such as spatial neural networks with geospatial principles [13, 15]. We hope that our work inspires further efforts to properly benchmark new methods on both synthetic and real-world data, thereby improving our understanding of the use cases and advantages of spatially-explicit models.

References

- 1 Zia U Ahmed, Kang Sun, Michael Shelly, and Lina Mu. Explainable artificial intelligence (XAI) for exploring spatial variability of lung and bronchus cancer (LBC) mortality rates in the contiguous USA. *Scientific Reports*, 11(1):1–15, 2021.
- 2 Colin M Beale, Jack J Lennon, Jon M Yearsley, Mark J Brewer, and David A Elston. Regression analysis of spatial data. *Ecology letters*, 13(2):246–264, 2010.
- 3 Chris Brunson, Stewart Fotheringham, and Martin Charlton. Geographically weighted regression. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 47(3):431–443, 1998.
- 4 Alexis Comber, Christopher Brunson, Martin Charlton, Guanpeng Dong, Richard Harris, Binbin Lu, Yihe Lü, Daisuke Murakami, Tomoki Nakaya, Yunqiang Wang, et al. A route map for successful applications of geographically weighted regression. *Geographical Analysis*, 55(1):155–178, 2023.
- 5 Matthew J Cracknell and Anya M Reading. Geological mapping using remote sensing data: a comparison of five machine learning algorithms, their response to variations in the spatial distribution of training data and the use of explicit spatial information. *Computers & Geosciences*, 63:22–33, 2014.
- 6 Sourav Das and Guy P Nason. Measuring the degree of non-stationarity of a time series. *Stat*, 5(1):295–305, 2016.
- 7 Zhenhong Du, Zhongyi Wang, Sensen Wu, Feng Zhang, and Renyi Liu. Geographically neural network weighted regression for the accurate estimation of spatial non-stationarity. *International Journal of Geographical Information Science*, 34(7):1353–1377, 2020.
- 8 Andrew O Finley. Comparing spatially-varying coefficients models for analysis of ecological data with non-stationary and anisotropic residual dependence. *Methods in Ecology and Evolution*, 2(2):143–154, 2011.
- 9 A Stewart Fotheringham, Chris Brunson, and Martin Charlton. *Geographically weighted regression: the analysis of spatially varying relationships*. John Wiley & Sons, Chichester, UK, 2003.
- 10 A Stewart Fotheringham, Wenbai Yang, and Wei Kang. Multiscale geographically weighted regression (MGWR). *Annals of the American Association of Geographers*, 107(6):1247–1265, 2017.
- 11 Stefanos Georganos, Tais Grippa, Assane Niang Gadiaga, Catherine Linard, Moritz Lennert, Sabine Vanhuyse, Nicholas Mboga, Eléonore Wolff, and Stamatis Kalogirou. Geographical Random Forests: a spatial extension of the Random Forest algorithm to address spatial heterogeneity in remote sensing and population modelling. *Geocarto International*, 36(2):121–136, 2021.
- 12 Daniel A Griffith and Yongwan Chun. Implementing Moran eigenvector spatial filtering for massively large georeferenced datasets. *International Journal of Geographical Information Science*, 33(9):1703–1717, 2019.
- 13 Julian Hagenauer and Marco Helbich. A geographically weighted artificial neural network. *International Journal of Geographical Information Science*, 36(2):215–235, 2022.
- 14 Tomislav Hengl, Madlene Nussbaum, Marvin N Wright, Gerard BM Heuvelink, and Benedikt Gräler. Random Forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. *PeerJ*, 6:e5518, 2018.

- 15 Konstantin Klemmer. *Improving neural networks for geospatial applications with geographic context embeddings*. PhD thesis, University of Warwick, Coventry, UK, 2022.
- 16 James LeSage. Spatial econometrics. In Charlie Karlsson, Martin Andersson, and Therese Norman, editors, *Handbook of research methods and applications in economic geography*, pages 23–40. Edward Elgar Publishing, Cheltenham, UK, 2015.
- 17 James P LeSage. A family of geographically weighted regression models. In Luc Anselin, Raymond J. G. M. Florax, and Sergio J. Rey, editors, *Advances in spatial econometrics*, pages 241–264. Springer, Berlin/Heidelberg, Germany, 2004.
- 18 Jin Li, Andrew D Heap, Anna Potter, and James J Daniell. Application of machine learning methods to spatial interpolation of environmental variables. *Environmental Modelling & Software*, 26(12):1647–1659, 2011.
- 19 Xiaojian Liu, Ourania Kounadi, and Raul Zurita-Milla. Incorporating spatial autocorrelation in machine learning models using spatial lag and eigenvector spatial filtering features. *ISPRS International Journal of Geo-Information*, 11(4):242, 2022.
- 20 Benjamin S Murphy. PyKrige: development of a Kriging toolkit for Python. In *American Geophysical Union Fall Meeting Abstracts*, volume 2014, pages H51K–0753, San Francisco, CA, USA, 2014.
- 21 J Keith Ord and Arthur Getis. Local spatial heteroscedasticity (LOSH). *The Annals of Regional Science*, 48:529–539, 2012.
- 22 R Kelley Pace and Ronald Barry. Sparse spatial autoregressions. *Statistics & Probability Letters*, 33(3):291–297, 1997.
- 23 F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
- 24 Martin Raubal. It’s the spatial data science, stupid! In *Spatial Data Science Symposium “Setting the Spatial Data Science Agenda”*, Santa Barbara, CA, US, 2019. Center for Spatial Studies at the University of California.
- 25 MGJ Rikken and RPG Van Rijn. *Soil pollution with heavy metals: in inquiry into spatial variation, cost of mapping and the risk evaluation of Copper, Cadmium, Lead and Zinc in the floodplains of the Meuse West of Stein, The Netherlands: field study report*. University of Utrecht, 1993.
- 26 Sebastian Santibanez, Tobia Lakes, and Marius Kloft. Performance analysis of some machine learning algorithms for regression under varying spatial autocorrelation. In *Proceedings of the 18th AGILE International Conference on Geographic Information Science*, pages 9–12, Lisbon, Portugal, 2015.
- 27 Fabián Santos, Valerie Graw, and Santiago Bonilla. A geographically weighted Random Forest approach for evaluate forest change drivers in the Northern Ecuadorian Amazon. *PloS one*, 14(12):e0226224, 2019.
- 28 Aleksandar Sekulić, Milan Kilibarda, Gerard B.M. Heuvelink, Mladen Nikolić, and Branislav Bajat. Random Forest spatial interpolation. *Remote Sensing*, 12(10):1687, 2020.
- 29 René Westerholt. Emphasising spatial structure in geosocial media data using spatial amplifier filtering. *Environment and Planning B: Urban Analytics and City Science*, 48(9):2842–2861, 2021.
- 30 René Westerholt, Bernd Resch, Franz-Benjamin Mocnik, and Dirk Hoffmeister. A statistical test on the local effects of spatially structured variance. *International Journal of Geographical Information Science*, 32(3):571–600, 2018.
- 31 Ryan Zhenqi Zhou, Yingjie Hu, Jill N Tirabassi, Yue Ma, and Zhen Xu. Deriving neighborhood-level diet and physical activity measurements from anonymized mobile phone location data for enhancing obesity estimation. *International Journal of Health Geographics*, 21(1):1–18, 2022.

Confidential, Decentralized Location-Based Data Services

Benjamin Adams  

Computer Science and Software Engineering, University of Canterbury, Christchurch, New Zealand

Abstract

There are many privacy risks when location data is collected and aggregated. We introduce the notion of using confidential smart contracts for building location-based decentralized applications that are privacy preserving. We describe a spatial library for smart contracts that run on Secret Network, a blockchain network that runs smart contracts in secure enclaves running in trusted execution environments. The library supports not only basic geometric operations but also cloaking and differential privacy mechanisms applied to spatial data stored in the contract.

2012 ACM Subject Classification Security and privacy → Human and societal aspects of security and privacy; Security and privacy → Economics of security and privacy

Keywords and phrases spatial data, privacy, smart contract, differential privacy

Digital Object Identifier 10.4230/LIPIcs.GIScience.2023.12

Category Short Paper

Supplementary Material *Software:* <https://github.com/darwinzer0/secret-data-tools>
archived at `swh:1:dir:aa0055da5ff292a15f6b44e5c54abf79fbfdab04`

1 Introduction

Mobile devices provide a number of opportunities to collect spatial data about human behavior, which can be used for data analytics and as training data for machine learning algorithms [10]. However, location data can also reveal extremely sensitive personal information and can easily be used in a harmful manner that violates an individual's privacy [3, 5]. Location data is not unique in this manner – centralized online platforms, such as large online social networks, have in recent years been critiqued for a wide-variety of ways that private user data is used and aggregated for commercial gain at the expense of user privacy. A recent trend toward building more decentralized applications that allow users greater control over their own data proposes to counteract this practice of gathering private data into centralized silos [8].

In this paper we present an approach to building decentralized location-based applications using confidential smart contracts that execute within hardware-encrypted environments. We demonstrate how we can use confidential smart contracts to implement spatial cloaking and to calculate summary statistics with global differential privacy on private location information without revealing individual information to others, such as a centralized server administrator. In addition, we explain how this model can allow for the creation of privacy-preserving location data marketplaces where data contributors can opt-in and control the amount of data they produce, control the level of spatial granularity, privacy thresholds, etc. on what data is made available, all while providing a mechanism for data producers to be directly paid for their contributions to the data set.

Figure 1 illustrates a sample scenario of computing over location data in a decentralized application that uses a confidential smart contract. Alice and Bob are users of mobile devices that allow them record their spatial location. They each execute a transaction on a confidential smart contract that stores their location (x, y) at a given timestamp on the chain



© Benjamin Adams;
licensed under Creative Commons License CC-BY 4.0

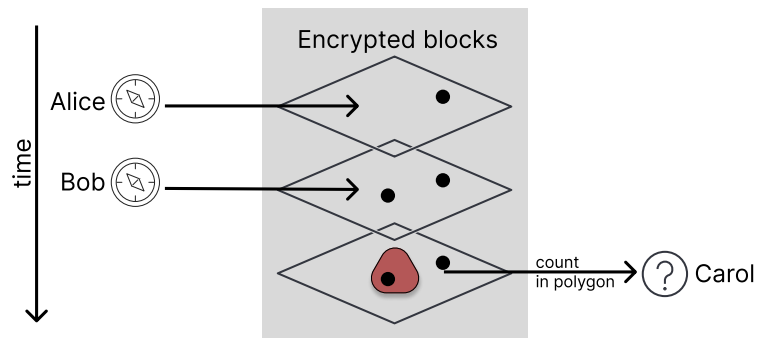
12th International Conference on Geographic Information Science (GIScience 2023).

Editors: Roger Beecham, Jed A. Long, Dianna Smith, Qunshan Zhao, and Sarah Wise; Article No. 12; pp. 12:1–12:6

Leibniz International Proceedings in Informatics



LIPIC Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany



■ **Figure 1** A simple scenario where private location points are stored on a blockchain and a third-party user makes a spatial query on the data without learning personal data.

in an encrypted format. They can query their own data points, but others, such as Carol, are only authorized to query for COUNT values within a given polygon. Permissions to access data are set by the logic of the contract and can be customized to suit any application. Furthermore, the permission system can be built to allow data owners to personalize settings, such as privacy thresholds or even charge for the use of their data before its acquisition.

2 Confidential smart contracts

Blockchains are essentially decentralized append-only databases [11]. They record transactional information as blocks of data that are generated by a peer-to-peer network of miners or validator nodes. The network operates using a consensus mechanism that ensures the history of the data (i.e., the chain) is immutable. In other words a bad actor cannot corrupt the chain and claim a transaction that did not occur and, for example, double spend some amount of currency recorded by the chain. The most well-known public blockchain, Bitcoin, uses a proof-of-work consensus mechanism which has an energy footprint that scales with the size of the network, but many other networks, including most that run smart contracts (described below), have long since shifted to proof-of-stake, which is far more energy efficient. By design all the data that is stored on a blockchain is visible, so it is a public database that allows one to examine every state of the chain, which means all transactions are by default public as well.

Smart contracts, first developed for the Ethereum network, are programs which can be run on blockchain networks [2]. Each node in the network runs a virtual machine that can execute code written in a turing-complete language. Ethereum uses the Solidity language running in the Ethereum Virtual Machine (EVM), while a number of newer chains run contracts that are written in general-purpose programming languages (e.g., Rust) and which are compiled to WebAssembly. Smart contracts are deterministic and modify the state of the data on the chain based on cryptographically-signed input messages that are sent to the network by client applications (often called decentralized apps or dapps). Smart contracts are referred to as trustless applications, because the logic of the contract is fixed and the consensus mechanism of the blockchain network ensures that messages sent to the chain will be interpreted in a fixed way based on the logic built into the program. This provides the ability to develop automated programs that allow users to conditionally transact on information without requiring a trusted third-party to verify that conditions have been met. Smart contracts have vastly increased the utility of blockchains leading to a wide-variety of applications in decentralized finance, digital art, and supply chain management.

Despite these many new kinds of dapps, the fact that all data and transactions that occur on most blockchains is completely transparent means that they are unsuitable for applications that require keeping information confidential, e.g. for user privacy. Some privacy blockchains utilize non-interactive zero knowledge proofs to provide transactional privacy built in, where proof of transactions from one account to another is recorded on the chain without revealing the amount of the transaction or which accounts were included [9]. However, the utility of zero knowledge proofs is limited to situations where there are only two parties involved. In other words, they are not capable of answering questions about data points that are individually private to a large number of different parties (e.g., calculating aggregate statistics).

However, there are a few blockchain networks actively developing more general-purpose confidential smart contract frameworks, where the internal state of smart contract execution as well as the data on chain is encrypted [14]. With confidential smart contracts you can create secure systems involving multiple clients that protect individually-supplied information while also computing over that information to provide outputs that are usable by other parties. Among three proposed methods: homomorphic encryption, secure multi-party computation (MPC), and trusted execution environments (TEEs), only the last (TEEs) is implemented in a working public blockchain. Fully homomorphic encryption is simply too slow to be a practical solution in current blockchain networks and secure MPC has also not been successfully implemented in a live network (although research remains active). TEEs however have been successfully integrated into live public blockchains since 2020 (with Secret Network¹ and more recently the Oasis Network²).

TEE-based smart contracts rely on using specific hardware chips where code executions can occur within a protected encrypted enclave [7]. For example, all the nodes in Secret Network are running on a set of Intel SGX chipsets. This means that the trust in the encryption of the network is based on trust that the hardware is secure. The upshot is that even the people who are running the computer nodes in the network cannot inspect the program state or data being used while the smart contract is executing, and all data written to the blocks are encrypted. In this paper we explore implementations of spatial algorithms using smart contracts running on Secret Network, at the moment the most mature network for developing confidential smart contracts. On Secret Network, contracts are written in Rust and compiled to WebAssembly before being uploaded to the chain. To date, most applications running on Secret Network are in the area of decentralized finance – this paper presents the first exploration into developing privacy-preserving location-based applications using the network.

3 Storing spatial data on chain

Developing on blockchain networks has limitations not found in normal programming environments. An important difference is that smart contracts cannot use floating point mathematical operations because they are not deterministic (different platforms implementing the IEEE 754 standard can output different results based on rounding), therefore there is a risk that different nodes in a network will be unable to come to consensus on floating point data written to chain. As a result to build a contract for operating on spatial data, e.g. points, lines, and polygons with x and y coordinates in a projected coordinate system,

¹ <https://scret.network/>

² <https://oasisprotocol.org/>

there are two options: 1) store locations on an integer grid using 128 bit or 256 bit integer values for the x and y coordinates, or 2) using a fixed-point representation where numbers are represented as rational numbers using software operations on integer data structures. In the first case, operations will be fast and with 128-bit and greater sized integers meaning that units can be expressed in micro-meters or smaller, which for all intents and purposes allows the same level of geographic precision as any floating point representation. However, the kinds of operations that can be executed will be limited. In the second case, we can utilize fixed-point math libraries with functions such as *sqrt* and transcendental functions, including *exp*, *sin*, *cos*, etc.

Despite not being able to do some mathematical operations, many common spatial queries are still possible on data represented on an integer grid. For example, point in polygon, line intersection, convex hull, range searching, and nearest neighbor are all possible. This is due that fact you can calculate length squared as the dot product of a vector with itself and calculate signed area for two vectors and infer turn relationships based on the sign. Some operations however do require a fixed-point representation, such as great circle distance or applying differential privacy techniques to return fuzzy statistics on the data set stored in the contract.

The other main limitation for storing data on the chain is that doing so can be expensive. Smart contracts meter the amount of computation and data written and read from the chain and charge a fee (i.e., a gas requirement) to manage the computational load on the network. Read-only queries are free, but anything that writes data to a new block on the chain will cost something as a factor of these. Paying a small upfront gas fee for storing personal location data on the chain might be acceptable to some users if it means they maintain control over it and in fact if they are able to directly commoditize their own data while also having access to location-based services.

4 Spatial library for confidential smart contracts

We have developed a `secret-data-tools` spatial package for creating spatial applications on Secret Network. Code is written in Rust and available on Github (see Supplementary Material). For both integer-grid and fixed-point the library provides a set of geometric primitive structs written in Rust for Points, LineSegments, and Polygons. It also includes a set of basic geometric query operations, such as point in polygon, that work for both integer and fixed-point representations.

Spatial cloaking is a method for masking location data points into a wider geographic region (or some minimum size) and in a manner that maintains a certain level of k-anonymity [6, 12]. Implementing spatial cloaking of data with the library functions is rather trivial. A contract can mask data into grid cells or other regions when producing answers to queries. Furthermore, because authorization of data access can be customized to any use case, users can e.g., choose to provide more granular data to specific individuals, categories of individuals, or applications.

For point data stored in fixed-point representation, the library provides an implementation of global ϵ -differential privacy using the Laplace mechanism [4]. Differential privacy adds noise to the result of a function, e.g. COUNT or AVERAGE, such that the result satisfies the constraint set by the privacy budget parameter, ϵ . Composed with the basic geometry operations, the library provides the capability to perform queries such as returning a fuzzy count of the number of spatial observations that fall within a polygon boundary, without revealing any information about individual data points.

Services that use differential privacy on location data, e.g. collected by mobile phone apps, will often utilize *local* differential privacy. With local differential privacy, noise is added to each individual data point prior to collection in a centralized database. This helps to maintain individual privacy, however, because noise is added to each observation, rather than only the final result, the overall accuracy of the data is degraded more quickly. However, using confidential smart contracts, because the data values are only visible to the contract itself when running in the protected enclave and no outside observer can view them, we can store direct observation values and implement global differential privacy without needing to trust a central data administrator.

A characteristic of differential privacy is that each query made on the database erodes the privacy budget. This happens because multiple queries on the same data can reveal the true value eventually, so for data stored on a blockchain we need to add additional limitations on the number of times that a query can be performed. Therefore, all differential privacy queries are implemented not as read-only queries but rather as read-write transactions that not only provide the answer but also update the remaining privacy budget on chain. The query will fail if not enough privacy budget remains for the query.

5 Toward building decentralized location-based applications

The `secret-data-tools` spatial package is a toolkit for building decentralized location-based applications on Secret Network. Using this library enables a number of different possibilities for data sharing and services. Allowing data contributors to set their own thresholds for acceptable data sharing can lead to fine-grained control over location data. Data sources need not be from individual, personal devices either. Other location-based data, such as from object tracking or transportation nodes, might require confidentiality for business processes.

If a user wishes to contribute their location-based information then they will have to pay to put the data on the chain. The amount paid depends on how much data and the parameters of the network – a small amount of data (e.g., an individual point observation) will cost a fraction of a penny, but larger amounts of data will quickly add up. Contracts can operate data marketplaces that require payment from data readers before releasing data, which can be directly paid to data contributors without the need for an intermediary. Furthermore, the privacy parameters of spatial queries (e.g., the size of masking regions or privacy budget) can be made to be user-settable.

A larger, practical concern is that providing a direct incentive for data sharing will also likely incentivize people to upload false information, given that GPS data can be easily spoofed. The use of confidential smart contracts for spatial data analysis is particularly well-suited to be paired with *proof-of-location* systems [1, 13]. New proof-of-location technologies in development, such as FOAM³ which uses networked LoRA devices to record location-based events, by necessity will require privacy-preserving mechanisms built-in prior to wide adoption. Currently they do not have that capacity, however. Confidential smart contracts provide one possible solution to incorporating privacy in proof-of-location systems, while at the same time proof-of-location can ensure fair decentralized marketplaces for spatial data.

Although we have primarily focused on examples of uploading individual data points, location-based data need not be stored as individual observations. Various methods of rolling up data are possible, which can be more efficient and result in lower gas charges for data contributors. In addition, there is the option of storing encrypted location data off the chain,

³ <https://foam.space/>

and storing only the decryption key on chain. In such a model there would be very little cost to the user, however many of the advantages of trustless computation on individual spatial data points from multiple contributors will be lost.

6 Conclusion

This paper presented a new approach for building decentralized programs that allow users to privately share location-based data using confidential smart contracts. We introduced an open-source library for Secret Network-based smart contracts, which includes basic geometry operations and can support spatial data cloaking and differentially private queries. There is more research that is required to fully evaluate efficacy of such tools to support different types of privacy-preserving location-based applications and data sharing platforms. In addition, a security analysis of potential side-channel attacks both in terms of the underlying blockchain technology, as well as based on inference from other data is warranted.

References

- 1 Michele Amoretti, Giacomo Brambilla, Francesco Medioli, and Francesco Zanichelli. Blockchain-based proof of location. In *2018 IEEE International Conference on Software Quality, Reliability and Security Companion (QRS-C)*, pages 146–153. IEEE, 2018.
- 2 Vitalik Buterin. A next-generation smart contract and decentralized application platform. Technical report, Ethereum Foundation, 2014.
- 3 Matt Duckham and Lars Kulik. Location privacy and location-aware computing. In *Dynamic and mobile GIS*, pages 63–80. CRC press, 2006.
- 4 Cynthia Dwork. Differential privacy. In *Automata, Languages and Programming: 33rd International Colloquium, ICALP 2006, Venice, Italy, July 10-14, 2006, Proceedings, Part II 33*, pages 1–12. Springer, 2006.
- 5 Hongbo Jiang, Jie Li, Ping Zhao, Fanzhi Zeng, Zhu Xiao, and Arun Iyengar. Location privacy-preserving mechanisms in location-based services: A comprehensive survey. *ACM Computing Surveys (CSUR)*, 54(1):1–36, 2021.
- 6 Mei-Po Kwan, Irene Casas, and Ben Schmitz. Protection of geoprivacy and accuracy of spatial information: How effective are geographical masks? *Cartographica: The International Journal for Geographic Information and Geovisualization*, 39(2):15–28, 2004.
- 7 Frank McKeen, Ilya Alexandrovich, Alex Berenzon, Carlos V Rozas, Hisham Shafi, Vedvyas Shanbhogue, and Uday R Savagaonkar. Innovative instructions and software model for isolated execution. *Hasp@ isca*, 10(1), 2013.
- 8 Christian Meurisch, Bekir Bayrak, and Max Mühlhäuser. Privacy-preserving AI services through data decentralization. In *Proceedings of The Web Conference 2020*, pages 190–200, 2020.
- 9 Xiaoqiang Sun, F Richard Yu, Peng Zhang, Zhiwei Sun, Weixin Xie, and Xiang Peng. A survey on zero-knowledge proof in blockchain. *IEEE network*, 35(4):198–205, 2021.
- 10 Eran Toch, Boaz Lerner, Eyal Ben-Zion, and Irad Ben-Gal. Analyzing large-scale human mobility data: a survey of machine learning methods and applications. *Knowledge and Information Systems*, 58:501–523, 2019.
- 11 Dylan Yaga, Peter Mell, Nik Roby, and Karen Scarfone. Blockchain technology overview. Technical report, National Institute of Standards and Technology, 2018.
- 12 Chengyang Zhang and Yan Huang. Cloaking locations for anonymous location based services: a hybrid approach. *GeoInformatica*, 13(2):159–182, 2009.
- 13 Pengxiang Zhao, Jesus Rodrigo Cedeno Jimenez, Maria Antonia Brovelli, and Ali Mansourian. Towards geospatial blockchain: A review of research on blockchain technology applied to geospatial data. *AGILE: GIScience Series*, 3:71, 2022.
- 14 Guy Zyskind, Oz Nathan, and Alex Pentland. Decentralizing privacy: Using blockchain to protect personal data. In *2015 IEEE Security and Privacy Workshops*, pages 180–184. IEEE, 2015.

Towards an Inclusive Urban Environment: A Participatory Approach for Collecting Spatial Accessibility Data in Zurich

Hoda Allahbakhshi ✉

Digital Society Initiative, University of Zürich, Switzerland
Department of Geography, University of Zürich, Switzerland

Abstract

The unprecedented rate of urbanization, along with the increase in the aging and disabled populations, bring about an increasing demand for public services and an inclusive urban environment that allows easy access to those facilities. Spatial Accessibility is a measure to assess how inclusive a city is and how easily public facilities can be reached from a specific location through movement in physical space or built environment.

A detailed geodata source of accessibility features is needed for reliable spatial accessibility assessment, such as sidewalk width, surface type, and incline. However, such data are not readily available due to the huge implication costs. Remote crowdsourcing data collection using Street View Imagery, so-called 'virtual audits' have been introduced as a valid, cost-efficient tool for accessibility data enrichment at scales compared to conventional methods because it enables involving more participants, saving more time by avoiding field visits and covering a larger area.

Therefore, in our pilot project, ZuriACT: Zurich Accessible CiTy, with the help of digital tools that allow for virtual inspections and measurements of accessibility features, we want to contribute to collecting and enriching accessibility information in the city of Zurich embedded in a citizen science project that will have both scientific and social impacts.

With the help of additional accessibility data produced in this project, the issues of an inclusive urban environment can be demonstrated by mapping the potential spatial inequalities in access to public facilities for disabled or restricted people in terms of mobility. Thus, this project provides helpful insight into implementing policy interventions for overcoming accessibility biases to ensure equitable services, particularly for people with disabilities, and contributes to creating an inclusive and sustainable urban environment. It goes without saying that an inclusive city is beneficial and impacts the quality of life of not only the population groups mentioned above but also the society at large.

2012 ACM Subject Classification Social and professional topics

Keywords and phrases Spatial accessibility, virtual audits, digital tools, mobility disability, citizen science, inclusive city, Zurich

Digital Object Identifier 10.4230/LIPIcs.GIScience.2023.13

Category Short Paper

Funding I thank the University of Zurich, the Digital Society Initiative, and Smart City Zurich for partially financing this research.

1 Introduction

It is projected that by 2050, about 70 percent of the world's population will live in urban environments, 15 percent of them will live with disabilities [10]. Moreover, the prediction shows that by 2050, the number of older people will reach 2 billion worldwide [12]. The unprecedented rate of urbanization, along with the increase in the aging and disabled populations, bring about an increasing demand for public services and access to those facilities. Depending on the infrastructure and design, the urban environment and physical



© Hoda Allahbakhshi;
licensed under Creative Commons License CC-BY 4.0

12th International Conference on Geographic Information Science (GIScience 2023).

Editors: Roger Beecham, Jed A. Long, Dianna Smith, Qunshan Zhao, and Sarah Wise; Article No. 13; pp. 13:1–13:6
Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

space can facilitate or impede the mobility and accessibility of the aforementioned population groups and consequently affect their active social and physical participation in society as well as their quality of life [11]. Besides, promoting accessible built environments such as easy-access buildings and barrier-free sidewalks is a key element for sustainable and inclusive cities and is of high societal importance. But how can we measure the inclusivity of a city? Spatial accessibility, traditionally defined as the “potential of opportunities for interaction” [7] and more concretely understood as how easily destinations such as services (e.g., medical centers, grocery stores, and banks), friends, or places of social interaction can be reached from a certain location through movement in physical space, is one of the measures which is also a crucial factor for supporting active and healthy aging and mobility.

A comprehensive geodata source of accessibility features is a prerequisite for accurate spatial accessibility assessment and therefore, urban inclusivity measurement. Examples of accessibility features, i.e., spatial features impeding or facilitating accessibility, are sidewalk inclination, crossings, and ramps. Accessibility features are crucial to disabled and mobility-restricted persons’ navigation and mobility. Still, they are usually not offered by commercial geodata providers [13] and are mostly not readily available in existing open-access geographic information databases such as Open Street Map (OSM) [5]. Moreover, existing routing services and digital maps, such as Google Maps and OSM, fail to provide practical guidance for the above-mentioned persons’ navigation due to the lack of relevant information on the needs of these user groups, which results in incomplete routing results or results that may not always reflect real-world conditions [6].

Different data collection methods have been applied to address this data gap issue and support the mobility of persons with disabilities (e.g., wheelchair users, visually impaired persons), which are traditionally conducted on the field applying on-field surveys [13], sensors (e.g., Global Positioning System) [16, 9], or a wide range of mobile applications (e.g., Vespucci [20], Go Map!! [4], and StreetComplete [21]). However, during the last few years, with the widespread use of the Internet, remote data collection using Street View Imagery (SVI), so-called ‘virtual audits’ has emerged as a valid alternative to expensive and time-consuming field visits [17]. The most famous and popular service for providing SVI worldwide is Google Street View (GSV) which is a basis for most virtual audits [14, 17, 15]. Virtual auditing allows users to remotely and manually measure and collect accessibility features by virtually walking in the city using the SVIs.

Collecting and maintaining detailed and up-to-date geographical information on accessibility is a considerably laborious, time-consuming, and expensive process. Hence, public partners usually avoid investing in such costly data collection [13]. Applying collaborative technologies such as citizen science helps address this challenge. Compared to the physical-based traditional methods, the virtual audit tools are easy-to-use, time and cost-efficient, and suitable for collaborative data collection, allowing the participants, particularly those who do not have the opportunity to do field visits for data collection, comfortably and safely collect detailed data at a larger scale wherever and whenever they want.

As mentioned earlier, publicly available geographical data sources such as OSM lack a considerable amount of accessibility information. For example, based on a recent study, only 2.3 percent of the OSM footpath data in Zurich include the inclination or steepness [3]. Besides, to the best of our knowledge, there has been no comprehensive geodatabase or data collection of accessibility information for the city of Zurich. Also, the city has launched no participatory data collection campaign in that regard.

Therefore, in our participatory project titled: ZuriACT (Zurich Accessible CiTy), for the first time, with the help of virtual audits, we want to take the initiative and contribute to providing a systematic and enriched dataset of the accessibility features in the selected study



■ **Figure 1** Study area: District 1 of the city of Zurich.

area of District 1 of the city of Zurich embedded in a citizen science project. District 1 of the city of Zurich (see Fig. 1) has been selected as the study area due to its topographical and geographic characteristics such as inclined streets, various public facilities (e.g., shopping streets, touristic attractions, main train station), a significant number of commuter populations, and centrality.

Also, we aim to contribute to a better understanding of spatial accessibility and its potential biases in the urban environment by providing an enriched accessibility database that can bring about essential information for reliable accessibility measurements, thereby equipping policymakers and urban planners with helpful insights into a more sustainable and inclusive environment for society, particularly persons with disabilities. Moreover, generating further new data can significantly contribute to scientific gaps in the accessibility analysis domain that have not been addressed so far due to a lack of appropriate, comprehensive open geographical data.

2 Method

2.1 Recruitment and Participants

A range of different marketing options will be used to inform citizens about the ZuriACT project idea, including the organization's websites (e.g., The City of Zurich, the organizations for people with disability, University of Zurich), e-newsletters, social media (e.g., LinkedIn, and Twitter), and distributing flyer in the study area. The communication and recruitment of citizens will also be conducted through the university webpage, where citizens can find further information about the project, as well as contact information and register for the study.

After screening the registered people based on the inclusion criteria, eligible participants will be contacted via email and asked to sign a consent form, including information about the study objectives and procedure, expected contribution, and participant compensation. Upon receipt of the informed consent, participants will be contacted to schedule meetings for different parts of the project, including focus group discussions and training sessions for data collection.

A total of 80–100 will be recruited for the study. As for eligibility criteria, participants must be cognitively healthy (assessed based on self-report) adults aged 18 and above, and belong to at least one of the population groups below:

13:4 Towards an Inclusive Urban Environment

1. Community-dwelling older adults aged 65 and above
2. Persons with situation mobility restrictions, such as parents with pushchairs
3. Mobility-disabled persons (e.g., wheelchair users)

2.2 Focus Groups

Our citizen science project focuses on co-creation, aiming to maximize the level of citizens' involvement in most or all stages of the project, including project design, data collection, and implementation [18]. To this end, we apply methods and tools for co-producing knowledge, such as focus group discussions [19]. In workshops, we bring together academics, citizens, and public and private partners to discuss the project's objective and contents, including initial ideation and data collection specifications. This helps gain experience from various perspectives and learn about the needs, knowledge, demand, and interests of different people involved in the project, laying the basis to adapt the project planning in a way that could be beneficial to all. An example of a similar initiative is the 'MIND Inclusion' project by Martínez-Molina et al. (2020), which focused on providing co-created accessible cognitive design tools for people with intellectual disabilities [8].

2.3 Spatial Accessibility Data Collection

We will use the Project Sidewalk tool for virtual audits by citizens. Project Sidewalk allows for collecting accessibility data at a large scale by anyone with an Internet connection and a web browser through GSV images. Examples of data that can be collected using this tool are "curb ramp", "missing curb ramp", "surface problems", "no sidewalk", "Obstacles in path", and "Others" [15]. Besides, it offers an excellent citizen science platform that allows laypersons to collect accessibility data comfortably via interactive onboarding and mission-based tasks. However, it lacks tools for collecting accessibility features that require measurements, such as sidewalk incline or width. Moreover, Project Sidewalk highly depends on GSV images which are sometimes outdated or do not cover the entire street network of the study area.

To address the data collection gaps using Project Sidewalk, we will use the Infra3D web-based tool [2], which is based on up-to-date and complete 3D SVI data "Strassenraum 3D" taken from car-mounted cameras from the entire city of Zurich developed by the Swiss company iNovitas [1] and also offers measurement tools. The "Strassenraum 3D" data has a higher and finer temporal resolution and spatial coverage than GSV and is updated every two years. The 3D images embedded in the infra3D web-based tool have been taken from an equipped vehicle and include all public roads (excluding motorways) and the whole tram network of Zurich city and selected cycle paths and squares. However, since Infra3D lacks a well-designed citizen science platform like Project Sidewalk, it might be challenging for laypeople and citizens to contribute to data collection using this tool. Therefore, to address this issue, we will involve persons with expertise in geographical data for virtual auditing using this tool.

During the data collection, through online forums or on-site social events, we ask participants to provide feedback or exchange information regarding their data collection experiences. The data collection will continue until obtaining the total coverage of the accessibility features in District 1. However, using the above-mentioned web tools, there will still be data gaps in the areas that were not reachable by the vehicle, such as stairs or narrow alleys or where GSVs are missing. Therefore, our virtual data collection will be limited to the areas traversed by the car or covered by GSV images using Infra3D or Project Sidewalk, respectively. To fill

this void, the accessibility features will have to be collected via on-site field surveys with the help of research assistants. This can happen by using the most commonly used smartphone apps for enriching and editing OSM data, such as “Vespucci” or “Go Map!!” which enables on-site accessibility data collection. The on-site data collection can also help verify the data derived remotely from virtual audits.

2.4 Discussion and Conclusion

In this project, we aim to contribute to filling the spatial accessibility data gap on sidewalks in Zurich with and for citizens by providing a systematic collection and enrichment of accessibility features utilizing digital tools, and virtual audits. The participatory design of this project involving citizens, researchers, and public partners allows for collecting and enriching a vast amount of detailed accessibility information across a larger geographical area during a shorter period, which not only contributes to considerable savings in time and resources compared to conventional data collection methods but also provides additional descriptive and spatial data to address crucial research and practical questions about the mobility and spatial accessibility of disabled people and how to realize an inclusive urban environment.

References

- 1 Strassenraum 3d, 2020. URL: <https://www.stadt-zuerich.ch/ted/de/index/taz/gestalten/strassenraum-3D.html>.
- 2 infra3d web-client, 2023. URL: <https://mailchi.mp/a2e2c47bc6cb/the-new-infra3d-release-315-is-now-live>.
- 3 Hoda Allahbakhshi and Robert Weibel. Assessing open street map spatial accessibility data quality at different geographical scales (working title). in prep.
- 4 B. Cogswell. Go map!!, 2018. URL: <https://github.com/bryceco/GoMap>.
- 5 OpenStreetMap contributors. Openstreetmap, 2004.
- 6 Jon E Froehlich, Anke M Brock, Anat Caspi, João Guerreiro, Kotaro Hara, Reuben Kirkham, Johannes Schöning, and Benjamin Tannert. Grand challenges in accessible maps. *interactions*, 26(2):78–81, 2019.
- 7 Walter G Hansen. How accessibility shapes land use. *Journal of the American Institute of planners*, 25(2):73–76, 1959.
- 8 Sandra Martínez-Molina, Michela Saretta, Andrea Giaretta, Alice Segalina, Simone Visentin, Rosa Almeida, Raquel Losada, Teresa Cid Bartolomé, Jorge Garcés-Ferrer, Valentina Conotter, et al. The mind inclusion app: assistive technology to foster the inclusion of persons with intellectual disabilities in their community. *Italian Journal of Special Education for Inclusion*, 8(2):291–300, 2020.
- 9 Hugh Matthews, Linda Beale, Phil Picton, and David Briggs. Modelling access with gis in urban systems (magus): capturing the experiences of wheelchair users. *Area*, 35(1):34–45, 2003.
- 10 United Nations. Good practices of accessible urban development, 2016.
- 11 World Health Organization. *Global age-friendly cities: A guide*. World Health Organization, 2007.
- 12 World Health Organization. Decade of healthy ageing 2020–2030, 2020.
- 13 Federico Prandi, Marco Soave, Federico Devigili, Raffaele De Amicis, and Alkis Astyakopoulos. Collaboratively collected geodata to support routing service for disabled people. In *Proceedings of the 11th international Symposium on Location-Based Services*, pages 67–79, 2014.

13:6 Towards an Inclusive Urban Environment

- 14 Manaswi Saha, Kotaro Hara, Soheil Behnezhad, Anthony Li, Michael Saugstad, Hanuma Maddali, Sage Chen, and Jon E Froehlich. A pilot deployment of an online tool for large-scale virtual auditing of urban accessibility. In *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility*, pages 305–306, 2017.
- 15 Manaswi Saha, Michael Saugstad, Hanuma Teja Maddali, Aileen Zeng, Ryan Holland, Steven Bower, Aditya Dash, Sage Chen, Anthony Li, Kotaro Hara, et al. Project sidewalk: A web-based crowdsourcing tool for collecting sidewalk accessibility data at scale. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2019.
- 16 Adam D Sobek and Harvey J Miller. U-access: a web-based system for routing pedestrians of differing abilities. *Journal of geographical systems*, 8:269–287, 2006.
- 17 Madeleine Steinmetz-Wood, Kabisha Velauthapillai, Grace O’Brien, and Nancy A Ross. Assessing the micro-scale environment using google street view: the virtual systematic tool for evaluating pedestrian streetscapes (virtual-steps). *BMC public health*, 19(1):1–11, 2019.
- 18 Eliseeva Tatiana, Olivia Höhener, David Michael Kretzer, Regina Lenart-Gansinieć, Anke Maatz, Mike Martin, Ursina Roffler, Susanne Tönsmann, Evgenia Tsianou, and Stefan Wiederkehr. *Practicing Citizen Science in Zurich: Handbook*. Citizen Science Center Zurich, 2021.
- 19 Peter Van Eeuwijk and Zuzanna Angehrn. How to... conduct a focus group discussion (fgd). methodological manual, 2017.
- 20 Vespucci. Osm editor, 2009. URL: <http://vespucci.io>.
- 21 T. Zwick. Streetcomplete, 2017. URL: <https://github.com/streetcomplete/StreetComplete>.

Development of a Semantic Segmentation Approach to Old-Map Comparison



Yves Annanias  

Image and Signal Processing Group, Leipzig University, Germany

Daniel Wiegrefe  

Image and Signal Processing Group, Leipzig University, Germany

Andreas Niekler  

Computational Humanities, Leipzig University, Germany

Marta Kuźma  

Faculty of History, University of Warsaw, Poland

Francis Harvey  

Leibniz Institute for Regional Geography, Leipzig, Germany

Faculty of History, University of Warsaw, Poland

Abstract

This paper describes an innovative computational approach for comparing old maps. Maps older than 20 years remain a vast treasure of geographic information in many parts of the world with potential applications in many environmental and social analyses, e.g., establishing road construction over the past 80 years or identifying settlement growth since the middle ages. Semantic segmentation has developed into a viable computational method for analysing old maps from previous centuries. It allows for the discrete identification of elements, e.g., lakes, forests, and roads, from cartographic sources and their computational modelling. Semantic segmentation uses convolutional neural networks to extract elements. With this technique, we create a computational approach to compare old maps systematically and efficiently.

2012 ACM Subject Classification Human-centered computing → Interactive systems and tools; Information systems → Geographic information systems

Keywords and phrases Geographic/Geospatial Visualization, Visual Knowledge Discovery, Cartographic Analysis

Digital Object Identifier 10.4230/LIPICs.GIScience.2023.14

Category Short Paper

Funding *Marta Kuźma and Francis Harvey:* The project is co-financed by the Polish National Agency for Academic Exchange within the NAWA Chair programme.

Yves Annanias: This research was supported by the Development Bank of Saxony (SAB) under Grant 100400221.

1 Introduction

Semantic segmentation is a computational method for analyzing old maps from previous centuries, allowing for discrete identification of elements like lakes, forests, and roads. This technique uses convolutional neural networks to extract the elements. The old maps used in this process contain valuable information, and comparing the elements they contain supports numerous environmental and social applications. Here, we present an innovative approach that allows us to compare multiple old maps. The paper considers the concepts and implementation and includes an assessment of the results of the new approach. Particularly challenging for this historical, geographical analysis are scale-related differences, distortions



© Yves Annanias, Daniel Wiegrefe, Andreas Niekler, Marta Kuźma, and Francis Harvey; licensed under Creative Commons License CC-BY 4.0

12th International Conference on Geographic Information Science (GIScience 2023).

Editors: Roger Beecham, Jed A. Long, Dianna Smith, Qunshan Zhao, and Sarah Wise; Article No. 14; pp. 14:1–14:6

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

of old map sheets, undocumented projection parameters and cartographic generalisation effects. The parametrisation of the semantic segmentation can take some geometric issues into account.

Our approach advances the handling of cartographic dimensions and will make systematic comparisons of collections of old maps possible and viable for the first time. For this, we construct a quadtree-based data structure that divides a map section and the features it contains into smaller and smaller sections, grouping them together. By visually displaying the levels of the quadtree as a heatmap, we then enable a more targeted comparison of features of the maps. Whereby the color coding highlights interesting map sections that may be of interest for a comparison. For the accurate and efficient modelling of the information from the old maps, we rely on a graph database that improves computational efficiencies of the cartographic element extraction and comparisons. In the paper, we document the modelling, processing and spatial visual comparisons of results of exemplary maps from the early and mid-twentieth centuries. The assessment of results points to challenges we are taking up in ongoing research.

2 Semantic Segmentation of Old Maps

Creating geographic information from old maps is an important source of data for many applications. For example, Uhl et al. [15] describe potentials for the over 200,000 topographic map sheets of the USGS map archive. While scanned versions of old maps are useful for wall hangings, screen savers and visual analysis by themselves, spatial analytical approaches frequently require additional processing to transform coordinate systems or features for specific project requirements. The transformation from raster to vector allows for other analytical operations that are well known from the development of GIS [5]. The cartographic modelling and geo-relational basis of those spatial analysis techniques is suitable for specific application and is limited by the computation complexity [14]. Database approaches are additionally advantageous when data can be optimised for requisite storage schemes and applications [4]. Machine learning approaches have for some years offered further computational improvements such as in [3] and are well-suited for the increasingly available large amounts of rasterised or vectorised geographic information.

2.1 Addressing cartographic challenges

Scale, distortions of old map sheets, undocumented projection parameters and cartographic generalisation effects are very significant challenges for any comparisons of old maps. Cartographic approaches, which stress graphic variables, concepts from cartographic design and features, build on traditional concepts of map representation that contemporary geographic information modelling approaches can never fully reconstruct [11]. The documented and archival information is usually very incomplete and research to gain insights involves much work and often only partial clarity. This can guide different modelling attempts. Often assumptions are made [13]. Old maps often are visually very insightful and intriguing documents of past geographical situations and relationships [17, 9]. Their accuracy is frequently limited and poses great challenges. In work using geographic information systems, the challenges are well known [8]. In cartography, research involves maps and specialised literature [6]. Their resolution is very time consuming. Integration of historical maps involves complicated and demanding data preparation and error mitigation [10]. We draw on these lessons and harness the capabilities of geographic information processing in our computational modelling. The computational approach in this research attempts to compare historical

maps, which computational approaches can greatly enhance help researchers move beyond the cartographic feature concept through the semantic segmentation process. The difference in terms stresses that the approach we describe here is information modelling approach to working with old maps.

3 Semantic segmentation for old map comparison

A critical part of working with old maps is determining the parameters for transforming digital raster scans of old maps into vector representations, suitable for CNN and normalisation of the coordinates for numerical pattern matching. Work on large-scale image analysis points the way for the approach we are developing. Therefore, we require a thorough documentation of processing steps and geometric attributes to allow for later assessments of comparison results including the identification of limitations arising from scale, distortions projections, or cartographic generalisation. Several researchers have addressed these issues [12, 15, 16, 18].

There are a variety of visualisations for geospatial and temporal data using a geographic information system (GIS). Andrienko et al. [1] provide a list of visualization-based techniques that allow the exploratory analysis of this kind of data. Since visual comparisons are essential in this task, we follow the guidelines of Gleicher [7]. In addition, as scalability also has an impact, we use the described strategy of *summarize somehow*. For this purpose, we rely for our approach on *explicit encoding*, whereby relationships between elements are visualised.

3.1 Process overview

Our approach follows the process presented in 2022 by Annanias et al. [2], but is simplified by limiting the area we consider in this pilot study, which focuses on a limited range of map element types and a small area. We adapted the color scheme, to fit to the new use case. The original version is used to aggregate data and show the distribution of that data over a larger area. With the limited map elements, it is now used to point out differences of similar elements. The parts of the the process are:

1. Implement shape comparisons between polygons in two maps using Hausdorff or Frechet distances and provide a system to support discovery and queries AND
2. Implement a GUI to compare multiple old maps by feature types or areas relying on visual opacity to support interactive visual inquiry.

3.2 Linking visual elements for further processing

The two parts of the process can be technically summarised as a five step sequence, whereby a quadtree-based data structure is created:

1. Determine the bounding box over all features, use it as the first parent cell.
2. Link all features to this parent cell.
3. Divide this cell into 4 equal parts (child cells).
4. Link all features from the parent cell to the child cell if they overlap with the child cell.
5. For each child cell, the process is repeated from step 3.

This process breaks the map image down and creates a quadtree, which consists of a grid of adjacent cells on each level (see Figure 1). As cells become smaller and therefore cover smaller areas of the map, the number of intersection calculations per cell becomes less. As a result, the test against the feature set of the parent cell becomes more computationally efficient. The number of cells, on the other hand, increases strongly. This information and all relationships are then stored efficiently and flexibly in a graph database. Each cell and



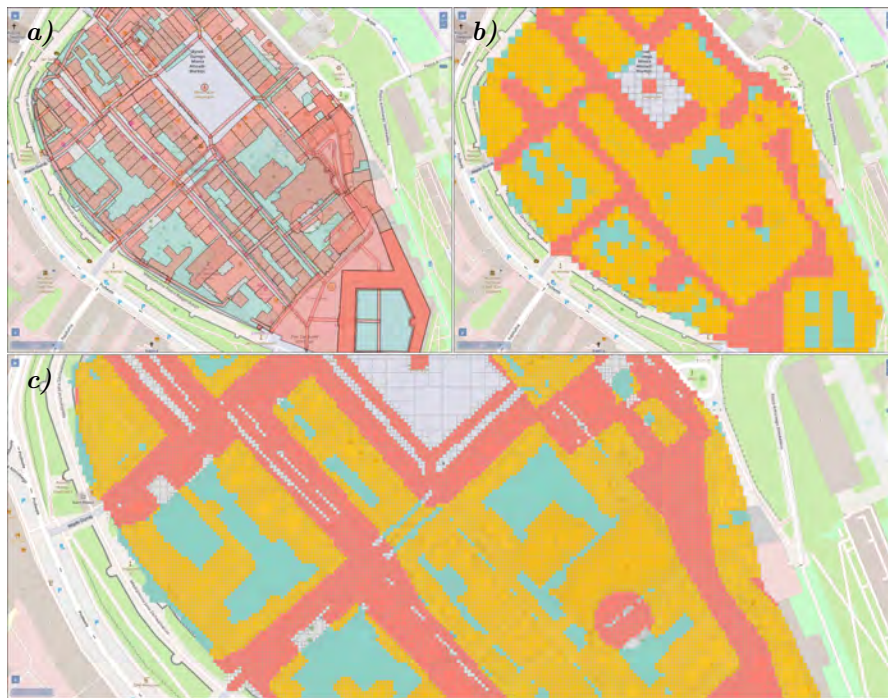
■ **Figure 1** From top to bottom: The initial map is divided progressively into equal parts, thus creating a quadtree with different resolution levels.

feature are represented by nodes connected by edges where the cell overlaps the feature. Cell nodes are also connected with each other by edges to represent the structure of the quadtree. In this way, the information for a grid level can be queried flexibly and the size of the overall graph becomes less important. At the most detailed grid resolution, only features that have a strong geographical adjacency are grouped together. At the lower levels of resolution, proximity in the quadtree is more diffuse and has a decreasing significance (e.g., a feature in one corner of a cell may have absolutely nothing to do with a feature in another corner of the same cell). Therefore, features that are too far apart no longer interact with each other. Because the process stops before reaching the next resolution level earlier, it avoids the extreme case, where each cell on the lowest level corresponds to only one piece of a feature (equivalent to perfect overlapping of two features) as the main task is not to find perfect overlaps of features. However, since offsets are also omitted and slight shifts of the features in relation to each other are no longer recorded, the process stops earlier after the 9th level. A cell on the lowest level has a resolution of about $1m^2$ in this study.

After the processing, each level of the quadtree can be used for visualisation. For this purpose, a level consisting of a grid of cells is represented as a heatmap in a GIS. So the heatmap is an aggregated representation of overlapping features (*summarize somehow*). Each cell of this heatmap is colored according to the relationships of the features that are linked to this cell (*explicit encoding*).

4 Results

The result is shown in Figure 2. Features from an old digitised map from 1941/1942 (blue) were used with OSM data (red), which are displayed superimposed in a). It is clearly visible that both feature categories overlap with each other. However, this overlap prevents us from seeing exactly how they overlap everywhere, as one obscures the other too much. So it is also important which category is displayed on top of which other. Similarly, if there are only small differences in detail, it is necessary to zoom in very close to see them, otherwise they may be overlooked. Figure 2 b) uses the same data, but uses a level from the quadtree and displays it as a heatmap (the previously created cells). The quadtree level with the highest resolution determines the color of a cell. Yellow cells indicate whether there are features from both categories within the cell. Cells of lower resolution levels inherit the color yellow if at least one of the four child cells is also marked yellow. This ensures that the features of both categories within a cell have a spatial proximity.



■ **Figure 2** a) Features of two maps are shown. b, c) Two resolution levels of the quadtree displayed as a heatmap (coarse to fine). Red (blue) cells contain only OSM (1941/1942) data features, and yellow cells contain at least one feature from both categories. White cells do not contain any features.

Using this visualization, it no longer matters which category is on top of the other, as the aggregated information for the cell is displayed. Similarly, subtle differences can no longer be overlooked. However, this is still a rough representation of the overlap and serves as a simple indication of areas of interest. This overview can be used, for example, to identify regions of interest in larger map segments. In doing so, a user can locate sub-areas through the larger grid cells, which can be viewed in detail by zooming and panning in the next step. c) shows the heatmap at a finer level of resolution. There is more detail here and it is easier to see where the features overlap and where they do not.

This allows the differences to be examined more closely without the visual clutter caused by the overlaps themselves. This representation thus serves as a starting point for the precise analysis of the shift of the categories towards each other. The comparison results support the visual comparison in a novel way that extends capabilities. Through an iteration of parameters, the resulting 'information spaces' extend canonical cartographic presentations to help researchers gain new insights into changes between two maps, for example assessing when a city's medieval walls were built up or torn down at various parts of a city.

5 Summary

In this paper, we present an innovative computational approach applied for comparing old maps. Showing good potential for historical research, the process has potential as well in other areas, e.g., assessments of urban development over the past 80 years or identifying ancient settlement growth. This preliminary result and other projects show that semantic segmentation is a viable computational method for the analysis of digitised old maps. This paper presents the computational process to compare old maps systematically and efficiently. Future research considers how to more fully automate the process and the comparisons.

References

- 1 Natalia Andrienko, Gennady Andrienko, and Peter Gatalsky. Exploratory spatio-temporal visualization: an analytical review. *Journal of Visual Languages & Computing*, 14(6):503–541, December 2003. doi:10.1016/S1045-926X(03)00046-6.
- 2 Yves Annanias, Dirk Zeckzer, Gerek Scheuermann, and Daniel Wiegrefe. An interactive decision support system for land reuse tasks. *IEEE Computer Graphics and Applications*, 42(6):72–83, 2022. doi:10.1109/MCG.2022.3175604.
- 3 Nicolas Audebert, Bertrand Le Saux, and Sébastien Lefèvre. Semantic segmentation of earth observation data using multimodal and multi-scale deep networks. In *Computer Vision–ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20–24, 2016, Revised Selected Papers, Part I 13*, pages 180–196. Springer, 2017.
- 4 Prafullata Auradkar, Tejas Prashanth, Suraj Aralihalli, Sreeniketh Pradeep Kumar, and Dinkar Sitaram. Performance tuning analysis of spatial operations on spatial hadoop cluster with ssd. *Procedia Computer Science*, 167:2253–2266, 2020.
- 5 Nicholas R. Chrisman. Design of geographic information systems based on social and cultural goals. *Photogrammetric Engineering and Remote Sensing*, 53(10):1367–1370, 1987.
- 6 Matthew Edney. *Cartography. The ideal and its history*. University of Chicago Press, Chicago, 2019.
- 7 M. Gleicher. Considerations for Visualizing Comparison. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):413–423, 2018. doi:10.1109/TVCG.2017.2744199.
- 8 Bernard Jenny, Helen Jenny, and Lorenz Hurni. Alte Karten als historische Quelle – Wie lässt sich die geometrische Genauigkeit des Karteninhalts abschätzen? In C. Koller and P. Jucker-Kupper, editors, *Karten, Kartographie und Geschichte – Von der Visualisierung der Macht zur Macht der Visualisierung / Cartes, cartographie et Histoire – De la visualisation du pouvoir au pouvoir de la visualisation*, pages 127–144. Chronos Verlag, Zürich, 2009.
- 9 Anne Kelly. Knowles and Amy Hillier. *Placing history : how maps, spatial data, and GIS are changing historical scholarship*. ESRI Press, Redlands, Calif., 2008.
- 10 J.B. Owens, May Yuan, M. Wachowicz, Vitit Kantabutra, E.A. Coppola, Daniel Ames, and Aldo Gangemi. Visualizing historical narratives: geographically-integrated history and dynamics gis. *library.wur.nl*, 2009.
- 11 Tomasz Panecki. Digital methods in cartographic source editing. *Digital scholarship in the humanities*, 36(3):682–697, 2020.
- 12 Rémi Guillaume Petitpierre, Frédéric Kaplan, and Isabella Di Lenardo. Generic semantic segmentation of historical maps. *CEUR Workshop Proceedings*, 2989(27):21. 228–248, 2021.
- 13 Stanisław Pietewicz. Analyse de l’exactitude de quelques cartes du xviiie, xviiiie et xixe siècle, couvrant les territoires de l’ancienne pologne. *Przegląd Geograficzny*, Special Issue for the XIX-th International Geographical Congress, Stockholm 1960:21–27, 1960.
- 14 Shashi Shekhar and Sanjay Chawla. *Spatial Databases. A Tour*. Pearson Education Inc., Upper Saddle River, NJ, 2003.
- 15 Johannes H Uhl, Stefan Leyk, Yao-Yi Chiang, Weiwei Duan, and Craig A Knoblock. Map archive mining: visual-analytical approaches to explore large historical map collections. *ISPRS international journal of geo-information*, 7(4):148–167, 2018.
- 16 Johannes H Uhl, Stefan Leyk, Yao-Yi Chiang, Weiwei Duan, and Craig A Knoblock. Automated extraction of human settlement patterns from historical topographic map series using weakly supervised convolutional neural networks. *IEEE Access*, 8:6978–6996, 2019.
- 17 M Wachowicz and JB Owens. The role of knowledge spaces in geographically-oriented history. *History and GIS*, 2013.
- 18 Kun Zheng, Ming Hui Xie, Jin Biao Zhang, Juan Xie, and Shu Hao Xia. A knowledge representation model based on the geographic spatiotemporal process. *International Journal of Geographical Information Science*, 36(4):674–691, 2022.

Why Is Greenwich so Common? Quantifying the Uniqueness of Multivariate Observations

Andrea Ballatore  

Department of Digital Humanities, King's College London, UK

Stefano Cavazzi  

Ordnance Survey, Southampton, UK

Abstract

The concept of uniqueness can play an important role when the assessment of an observation's distinctiveness is essential. This article introduces a distance-based uniqueness measure that quantifies the relative rarity or commonness of a multi-variate observation within a dataset. Unique observations exhibit rare *combinations* of values, and not necessarily extreme values. Taking a cognitive psychological perspective, our measure defines uniqueness as the sum of distances between a target observation and all other observations. After presenting the measure u and its corresponding standardised version u_z , we propose a method to calculate a p value through a probability density function. We then demonstrate the measure's behaviour in a case study on the uniqueness of Greater London boroughs, based on real-world socioeconomic variables. This initial investigation indicates that u can support exploratory data analysis.

2012 ACM Subject Classification Mathematics of computing → Multivariate statistics; Information systems → Geographic information systems

Keywords and phrases uniqueness, distinctiveness, similarity, outlier detection, multivariate data

Digital Object Identifier 10.4230/LIPIcs.GIScience.2023.15

Category Short Paper

Supplementary Material *Software (R Code):*

<https://github.com/andrea-ballatore/calculating-uniqueness>

Funding This work was supported by Ordnance Survey.

1 Introduction

The identification of similar observations is a well-studied problem in data science [2]. All forms of clustering rely on some form of similarity assessment [7], and so does data deduplication. Online platforms relentlessly search for similar products and similar users to drive engagement and sales. Geographic concepts can be compared and grouped by similarity and relatedness [1]. Geo-demographic classifications group similar areas based on socioeconomic characteristics. Similarity enables the identification of points representing points that appear close in a multi-dimensional vector space [7].

But what about uniqueness, one of the similarity's less known siblings? In many domains, the degree to which an object is unique is crucial to assess its value. The distinctiveness of artworks is carefully studied by scholars to ascertain the originality of painters, musicians, and writers. Unique cities, landscapes, and heritage assets are praised in the rhetoric of tourism marketing [9]. In the natural sciences, uniqueness is useful to define physical or chemical properties, genetic or molecular characteristics, or ecological traits that distinguish an individual from all others. Unique fingerprints, faces, irises, and DNA sequences enable ubiquitous applications in cybersecurity and forensic science. The cognate concept of "distinctiveness" is used in biology to explore the taxonomic structure of species [4]. In



© Andrea Ballatore and Stefano Cavazzi;
licensed under Creative Commons License CC-BY 4.0

12th International Conference on Geographic Information Science (GIScience 2023).

Editors: Roger Beecham, Jed A. Long, Dianna Smith, Qunshan Zhao, and Sarah Wise; Article No. 15; pp. 15:1–15:6

Leibniz International Proceedings in Informatics



LIPIC Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

recommender systems, it has been deployed to assess the typicality of user preferences [8]. The uniqueness of observations has been occasionally operationalised to support the interpretation and filtering of multi-dimensional data [5]. In its infrequent appearances in the scientific literature, this concept is largely left unexamined.

A relevant and intensely investigated problem is that of outlier detection, which consists of identifying unusual observations that might indicate infrequent but interesting events (e.g., fraudulent bank transactions) or measurement errors. Outliers can be found with distance-, clustering-, density-, ensemble-, and learning-based methods, with varying levels of success and robustness [10]. In a multivariate context, outliers are *rare combinations of values*. The values of each variable might not be extreme, but the combination appears relatively far from the others. The Mahalanobis distance is very useful for finding outliers in multidimensional datasets. While linked to outlier detection, our objective is the quantification of uniqueness as a facet of observations for classification and exploratory data analysis.

To support the exploration and recommendation of walking routes [3], we devised a distance-based uniqueness measure that can quantify how relatively rare (or common) an observation is in a dataset. In the spirit of classic ecological indices that have been in vogue for 40 years [11], we devise a simple, general, and easily interpretable measure that can be applied to many contexts. We ground our notion of uniqueness into a cognitive psychological perspective that defines the “distinctiveness of stimuli” as “the sum of the differences between the stimulus and all other stimuli in the group” [6, p. 16]. Hence, the more a multi-variate observation is different to all others, the more it is unique. The relative rarity of observations can act as an informative feature in machine learning methods and can support user interaction and data interpretation. Our measure u is described in the remainder of this article. Its behaviour is illustrated in a case study on London boroughs.

2 A uniqueness measure

Univariate uniqueness

In its simplest form, uniqueness can be thought of as $1 - p$, where p is the probability of encountering a particular observation from random extractions from a set. For example, let us consider the percentages of land cover categories of the UK territory: *farmland* 56.7%, *natural* 34.9%, *green urban* 2.5%, *built* 5.9%.¹ Taking this probabilistic view, the rarest category we would encounter by selecting a random area is green urban, corresponding to $u = .98$, and the most common (least unique) category is farmland ($u = .43$). This is conceptually linked to the idea of surprise – a less likely outcome is more surprising.

To make u more interpretable, we can calculate the corresponding z scores, relating uniqueness to the deviation from the normal distribution – with an assumption of normality that might not hold. If all types of observation occur at the same probability, it is not possible to meaningfully calculate uniqueness (z is null). Otherwise, common observations have negative z scores, and rare ones are positive: *farmland* has $z = -1.24$, while *green urban* has the highest value, with $z = .88$.

Multivariate uniqueness

As part of our efforts to support the exploration of large datasets [3], we developed a uniqueness measure that can handle multivariate observations. Considering a set of observations, the frequency of a given multi-variate configuration is correlated to uniqueness as rare observations

¹ <https://www.eea.europa.eu/publications/CORO-landcover>

are more unique than common ones. From a statistical standpoint, the assessment of uniqueness is also analogous with outlier detection, in which observations at the extreme of a distribution can be identified as of particular interest or as the result of measurement errors.

Our uniqueness index u between a multi-variate observation is calculated as the sum of the similarity of the observation with all other observations in the group S , ranging from rare to very common. Formally, given a set of observations S , the uniqueness score u of an observation $a \in S$ is defined as:

$$u(a, S) = \sum_{i=1}^{|S|} d(a, a_i), \quad a \neq a_i, \quad d \geq 0, \quad u \geq 0, \quad S = \{a_1 \cdots a_m\}, \quad u_z = \frac{u - \hat{u}}{\sigma(u)}$$

where d is an n -dimensional distance function. Different functions, such as Euclidean, Manhattan, or Mahalanobis, might produce radically different u . This measure is also sensitive to the particular structure of the data and to the selected variables. The scores are then standardised as u_z as z scores, where \hat{u} is the mean u and $\sigma(u)$ the standard deviation. The u_z scores are more interpretable than u , as they embed a measure of distance from the dominant clusters in the data. In other words, the index allows comparing observations on a spectrum ranging from very common (low values) to very rare (high values of u_z). An intuitive interpretation of these scores relates to the distance from cluster centres in the data space: Central data is common, and peripheral is rare.

In order to provide a measure of statistical significance, we calculate a p value for each standardised $z(u)$ using a probability density function of a normal distribution, defined as $f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}$, with \hat{u}_z as the mean μ and σ_{u_z} as σ . This approach captures the extent to which a value of u_z is greater than expected, with lower p for rarer cases. The underlying assumption is that u scores have an approximately normal distribution. This is a simplification of real-world data that might exhibit very different distributions of u and should be adjusted to specific contexts, but it is useful to develop our measure.

The behaviour of u_z and p values was tested on synthetic datasets of multivariate data, with a Monte Carlo approach, generating random high-dimensional datasets with different distributions and calculating u_z and corresponding p values, considering uniform, normal, and clustered distributions with two and three large clusters. In this initial empirical investigation, the observation-by-attribute matrices showed that the distribution of uniqueness scores remains fairly stable across different matrix sizes and across different distributions, although low p values are more frequently produced with clustered data. For example, considering 1440 matrices, on average, a normal distribution produced 0.54% of p smaller than .001 and 89% at $p > .1$, which makes intuitive sense.

Interpreting uniqueness

In a focus group we conducted at Ordnance Survey [3], we discussed the semantic interpretation of these scores with stakeholders. The uniqueness measure u_z was presented with walking routes as items to score, based on a number of attributes to identify unusual routes. The term “uniqueness” was considered semantically clearer than “distinctiveness.” From a cognitive perspective, participants expressed a preference for a categorical classification as opposed to both scores and ranks. Less agreement was found on the specific categories to use. The terms discussed included “common”, “rare”, and “typical” with modifiers “very” and “extremely”. It was noted that terms should not have positive or negative connotations, devaluing common items as uninteresting or valuing rare items that might be uncommon for good reasons – an unusual walking route around a landfill. Moreover, the participants highlighted the importance of showing the discriminant attributes along with the scores.

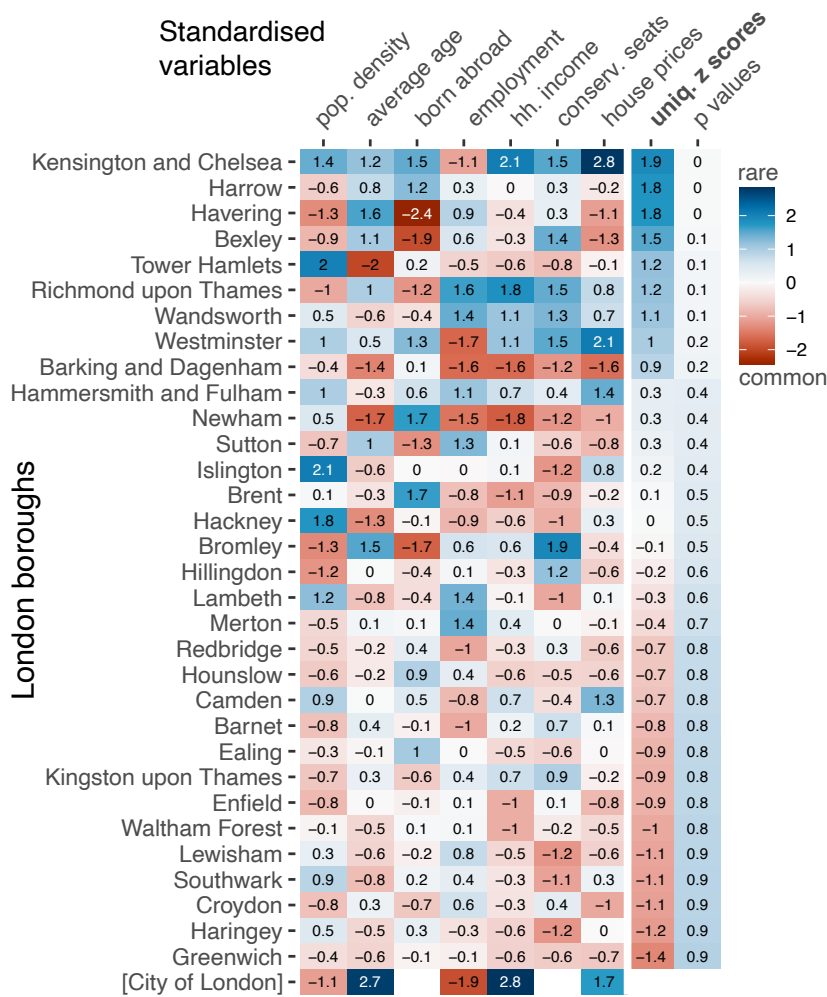
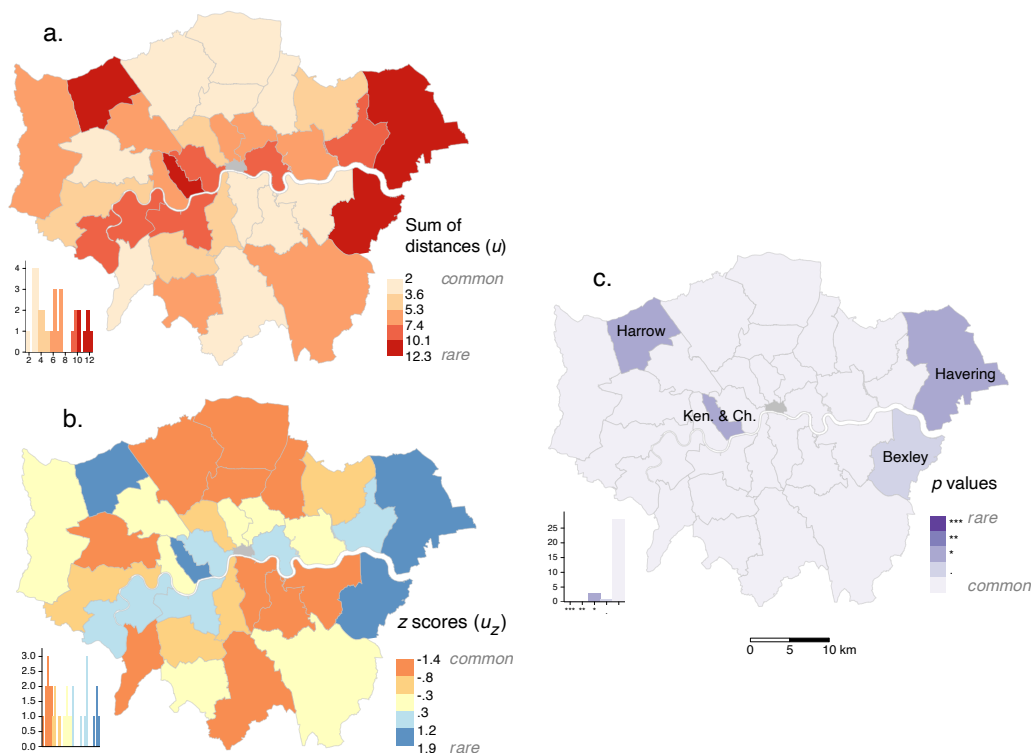


Figure 1 Uniqueness of 32 London boroughs with respect to seven socioeconomic variables, excluding City of London, calculated with the Mahalanobis distance. Variables were centred and standardised. The rows are ordered by the uniqueness z scores (u_z) from the rarest (Kensington and Chelsea) to the most common (Greenwich). Data source: London borough profiles, 2015.

As a result of this process, we defined five uniqueness levels based on p values as follows: ($p = 0$) *very rare* (.001) *rare* (.01) *intermediate* (.05) *common* (.1) *very common* (1). For example, $p = .006$ would be classified as rare. While such classifications are inevitably domain-dependent, these bins appear easily interpretable as they segment the scores at common p value thresholds.

3 The uniqueness of London boroughs

As an exploratory case study, we consider the boroughs of Greater London, a familiar, well-understood geography described through a set of socioeconomic variables. The seven selected variables include population density, average age, percentage of residents born abroad, percentage of employed residents, household income, Conservative seats, and median



■ **Figure 2** Uniqueness of 32 London boroughs with respect to seven socioeconomic variables, excluding City of London, calculated with Mahalanobis distance. (a) Sum of distances u ; (b) Uniqueness z scores (u_z); (c) p values with thresholds . ($p < .1$), * ($p < .05$), ** ($p < .01$), *** ($p < .001$). Bins for (a) and (b) were produced with Jenks. Data source: London borough profiles, 2015. Projection: British National Grid (EPSG:27700).

house price.² Highly correlated variables such as Labour seats ($\rho < -.7$ or $\rho > .7$) were removed to avoid obvious collinearity issues. Using our R implementation, we performed calculations of u , u_z , and p values for all boroughs, except for the City of London, which had several missing values. Among many possible options, the Mahalanobis distance measure was selected as it inherently accounts for scale invariance. Figure 1 presents the standardised matrix used for the calculation, along with the corresponding u_z and p values.

Boroughs exhibiting the highest u values are characterized by unexpected combinations of variables. According to our calculation method, three boroughs stand out as significantly rare, with uniqueness scores ($p < .05$). Kensington and Chelsea demonstrate exceptionally high house prices and household income, but relatively low levels of employment. Harrow, on the other hand, is intriguing as its variables are not extreme individually, but noticeably distinct from all other boroughs as a whole. Lastly, Havering appears relatively unique due to its ageing population and predominantly UK-born residents. In contrast, Croydon, Haringey, and Greenwich, located towards the bottom of the matrix, exhibit more central positions in the data, making them more representative of London as a whole. Greenwich, at least based on these variables, emerges as a very typical – and therefore common in our parlance – borough of Greater London. Figure 2 displays maps illustrating the spatial distribution of u , u_z , and p values. Visually, the three rare boroughs do not exhibit clustering.

² Data source: London Borough Profiles and Atlas, Greater London Authority (GLA), 2015.

The results from this analysis indicate that our measure, u , shows promise in quantifying the uniqueness of multivariate observations. However, further empirical testing with both real-world and synthetic data is necessary to assess the stability and interpretability of this uniqueness measure across different domains. A noteworthy characteristic of u is its weak correlation with any of the seven variables (the strongest correlation being $\rho = .39$). This suggests that the measure is capturing a latent dimension of the data, i.e. the distribution of uniqueness of these observations, revealing this facet of the data for further analysis.

In conclusion, further empirical testing is necessary to evaluate the stability and cognitive plausibility of this uniqueness measure across domains. Comparing different distance measures and methods for calculating p values is crucial to assess u 's sensitivity to minor data variations. The operationalisation of uniqueness might support meaningful analyses of why some places, cultural artefacts, human behaviours, and natural environments emerge as unique from a vast sea of sameness.

References

- 1 Andrea Ballatore, Michela Bertolotto, and David C. Wilson. An evaluative baseline for geo-semantic relatedness and similarity. *GeoInformatica*, 18:747–767, 2014. doi:10.1007/s10707-013-0197-8.
- 2 Andrea Ballatore, Michela Bertolotto, and David C Wilson. A structural-lexical measure of semantic similarity for geo-knowledge graphs. *ISPRS International Journal of Geo-Information*, 4(2):471–492, 2015. doi:10.3390/ijgi4020471.
- 3 Andrea Ballatore, Stefano Cavazzi, and Jeremy Morley. The context of outdoor walking: A classification of user-generated routes. *The Geographical Journal*, 2023. doi:10.1111/geoj.12511.
- 4 K Robert Clarke and Richard M Warwick. A taxonomic distinctness index and its statistical properties. *Journal of Applied Ecology*, 35(4):523–531, 1998.
- 5 Pamela J Ludford, Dan Cosley, Dan Frankowski, and Loren Terveen. Think different: increasing online community participation using uniqueness and group dissimilarity. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 631–638, New York, 2004. ACM.
- 6 Bennet B Murdock Jr. The distinctiveness of stimuli. *Psychological review*, 67(1):16–31, 1960.
- 7 Yazhou Ren, Jingyu Pu, Zhimeng Yang, Jie Xu, Guofeng Li, Xiaorong Pu, Philip S Yu, and Lifang He. Deep clustering: A comprehensive survey. *arXiv preprint*, 2022. arXiv:2210.04142.
- 8 Haggai Roitman, David Carmel, Yosi Mass, and Iris Eiron. Modeling the uniqueness of the user preferences for recommendation systems. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 777–780, New York, 2013. ACM.
- 9 Jonathan Schifferes. Mapping heritage. *RSA Journal*, 161(5563):10–13, 2015.
- 10 Hongzhi Wang, Mohamed Jaward Bah, and Mohamed Hammad. Progress in outlier detection techniques: A survey. *IEEE Access*, 7:107964–108000, 2019.
- 11 HG Washington. Diversity, biotic and similarity indices: A review with special relevance to aquatic ecosystems. *Water Research*, 18(6):653–694, 1984.

When Everything Is “Nearby”: How Airbnb Listings in New York City Exaggerate Proximity

Mikael Brunila  

Platial Analysis Lab, Department of Geography, McGill University, Montréal, Canada
Urban Politics & Governance Lab, School of Urban Planning, McGill University, Montréal, Canada

Priyanka Verma  

Platial Analysis Lab, Department of Geography, McGill University, Montréal, Canada

Grant McKenzie  

Platial Analysis Lab, Department of Geography, McGill University, Montréal, Canada

Abstract

In recent years, the emergence and rapid growth of short-term rental (STR) markets has exerted considerable influence on real estate in most large cities across the world. Central location and transit access are two primary factors associated with the prevalence and expansion of STRs, including Airbnbs. Nevertheless, perhaps due to methodological challenges, no research has addressed how location and proximity are represented in the titles and descriptions of STRs. In this paper, we introduce a new methodological pipeline to extract spatial relations from text and show that expressions of distance in STR listings can indeed be quantified and measured against real-world distances. We then comparatively analyze Airbnb reviews (written by guests) and listings (written by hosts) from New York City in order to demonstrate systematically how listings exaggerate proximity compared to reviews. Moreover, we discover spatial patterns to these differences that warrant further investigation.

2012 ACM Subject Classification Information systems → Geographic information systems; Information systems → Information extraction

Keywords and phrases spatial proximity, distance estimation, information extraction, named entity recognition, short-term rentals

Digital Object Identifier 10.4230/LIPIcs.GIScience.2023.16

Category Short Paper

Funding Mikael Brunila: Kone Foundation

Priyanka Verma: Social Sciences and Humanities Research Council of Canada

1 Introduction

Over the past decade, the short-term rental (STR) market has expanded in most large cities across the world. While STRs provide new economic opportunities for some, they also contribute to harmful processes such as gentrification and displacement through the removal of affordable units from the rental market [18, 1]. Airbnb in particular has become synonymous with a certain kind of gentrification, whereby conveniently located working-class and non-white neighborhoods are marketed as sites of consumption, leisure, and urban authenticity for an upwardly mobile class of white-collar professionals.

As in any market, Airbnb hosts need to communicate important information about location and other characteristics of their units to potential guests. Drawing on ideas from the interactional sociology of Ervin Goffman [6], ethnographers have likened Airbnb listings to front-stage performances whereby hosts deploy various means to manage the impressions guests will have of a listing [16]. Because real estate listings are fundamentally located *somewhere*, location figures strongly into the repertoire of *distinctions* [2] hosts can make



© Mikael Brunila, Priyanka Verma, and Grant McKenzie;
licensed under Creative Commons License CC-BY 4.0

12th International Conference on Geographic Information Science (GIScience 2023).

Editors: Roger Beecham, Jed A. Long, Dianna Smith, Qunshan Zhao, and Sarah Wise; Article No. 16; pp. 16:1–16:8

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

vis-a-vis other listings. Indeed, recent research has demonstrated that location is one of the key determinants of both the average price per night and average monthly revenue of units listed on Airbnb [4]. Nevertheless, this line of inquiry has not yet been extended to the “cognitive maps” [8, 10] which translate our experience of the city into mental representations thereof and, in the context of Airbnb, which encode the relationship between residence and place for people who typically reside elsewhere. By making claims about what is “nearby”, “only 10 minutes away” or “within walking distance”, Airbnb hosts situate their properties within the ensemble of a city’s structures and relations, including not only spatial and semiotic [10] but also ideological [8].

A wide range of work from various fields has established that our conception of what is “nearby” varies with a number of factors: larger objects tend to be considered closer than smaller ones, distances will be estimated differently depending on familiarity and activity, and so on (for an overview, see [5]). By comparing expressions of distance in listings and reviews, we can grasp how socio-economic incentives shape the production of spatial representations in discourse. While there is anecdotal evidence of the exaggeration of distance in the context of real estate advertisements [13], our paper provides a first glimpse into how these dynamics systematically unfold in a much larger dataset and in the setting of STRs. Furthermore, while others have presented models for extracting vague spatial descriptions [3, 5] as well as for assessing the linguistic distribution of concepts like “near,” we provide a sociological control variable by contrasting A) listing *descriptions* with B) listing *reviews* associated with the same locations. While listing descriptions are arguably written as a profit-motivated performance for the STR market, reviewers have different motives.

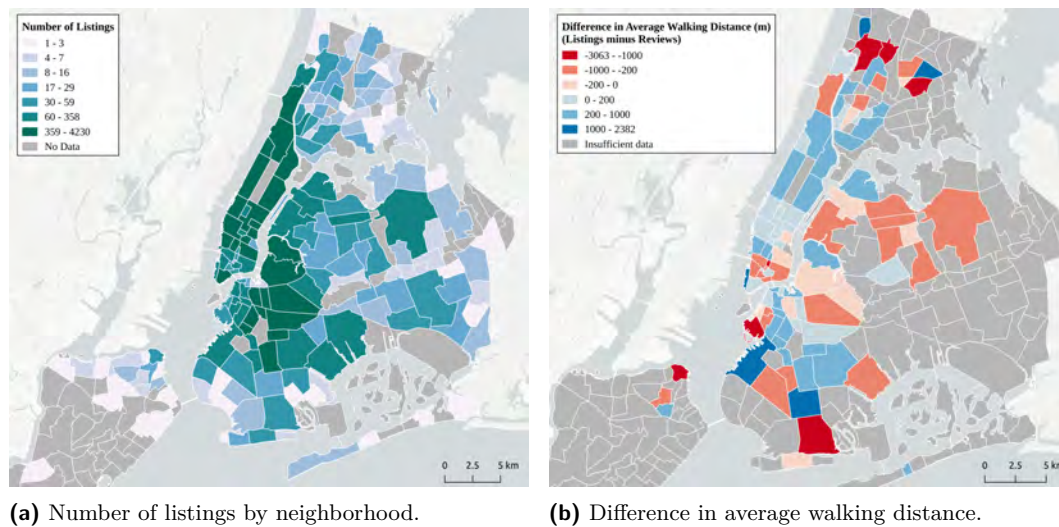
In this paper, we introduce a new methodological pipeline to extract spatial relations from text and show that expressions of distance in STR listings can indeed be quantified and measured against real-world distances. With this data, we demonstrate differences in the use of terms such as “nearby” and “walking distance” across listings and reviews. We do the same for a range of toponymic categories including parks (e.g., “Central Park”), tourist attractions (e.g., “Empire State Building”), and schools (e.g., “Columbia University”). Specifically, this short paper presents preliminary work addressing the following four research questions (RQ):

- RQ1 Can qualitative distance measures, such as *nearby* or *walking distance*, be quantified in STR listings?
- RQ2 Do quantified distance measures in STR listings accurately reflect real-world distances?
- RQ3 On average, do these distances vary between listing descriptions and reviews?
- RQ4 How do the above measures vary across neighborhoods in New York City (NYC)?

2 Data and Methods

The data for this paper cover all active Airbnb listings and their associated reviews for NYC in August 2019. All data were purchased from the non-profit group Inside Airbnb¹. The data contain a total of 47,440 listings and 995,665 reviews. To make data processing more feasible, we take a sample from the latter, giving us 168,533 reviews for an average 3.55 reviews per listing (even processing this sample takes a full day). Each listing includes its title, description, and geographic coordinates. The listings are highly unevenly distributed across the city, as can be seen in Figure 1a.

¹ <http://insideairbnb.com/>



■ **Figure 1** Cartographic representations of the Airbnb listing data. Raw count of listings per neighborhood shown in (a) and difference in average walking distance between listings and reviews shown in (b).

To extract geographical entities from the data, we manually annotated a spatially weighted random sample of 1,517 listings and 967 reviews using the annotation platform Prodigy² (for details on our sampling strategy, see Appendix A.1). The annotation was done by five academic annotators, including all the authors of this paper. This was effectively a named-entity recognition (NER) task, where the named entities were beyond the scope of existing general-purpose NER datasets. Annotators had 14 labels to choose between, which can be seen in Table 1 in the Appendix. For the purposes of this paper, the key labels are: (1) “Spatio-Temporal Entity” (STE) reflecting any relation between two locations, such as “15 minutes walk to” or “nearby,” and (2) various toponyms ranging from tourist attractions to schools. Labels were chosen from an initial set suggested by Cadorel et al. [3] but adjusted and extended to fit our specific dataset and framework.

After annotating the data, we fit three models with DistilBERT embeddings [17]. Out of these, a model with a Conditional Random Fields (CRF) [9] classification layer performed the best, with an overall F1-score of 0.756, with a plain DistilBERT model achieving comparable results with an F1-score of 0.752. To make our work more reproducible, we use this latter model, even if it is technically slightly worse. To connect STEs with relevant toponyms, we use a combination of dependency parsing and graph partitioning: Each STE is associated with the set of toponyms that are among its immediate dependents (for all the models and other details, see the Appendix and Table 1).

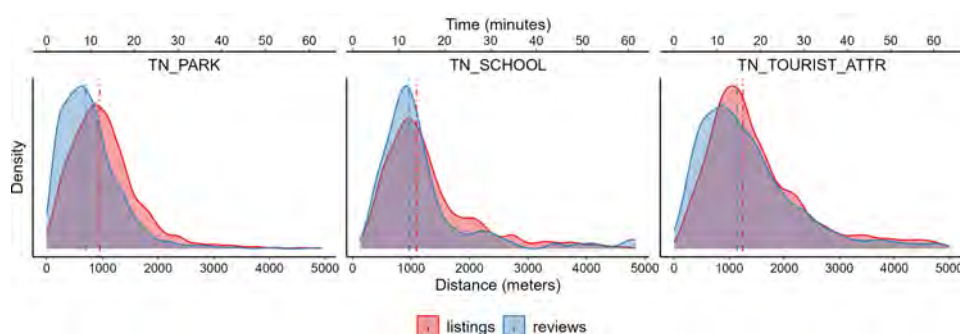
To address RQ1, we geocoded entities using Google’s Geocoding API.³ The geocoder provides coordinates for the centroid of each entity location. This poses a challenge for larger parks such as Central Park, which expands across an area of 3.41 km^2 . To address this issue, we calculated the point within the park closest to the Airbnb coordinates and used this as the final entity coordinates. We then used Open Source Routing Machine⁴ to

² <https://prodi.gy/>

³ <https://developers.google.com/maps/documentation/geocoding/overview>

⁴ <https://project-osrm.org/>

16:4 When Everything Is “Nearby”



■ **Figure 2** The distribution of distances (bottom x -axis) and walk times (top x -axis) for listings and reviews respectively. Listings consistently under-represent distance compared with reviews.

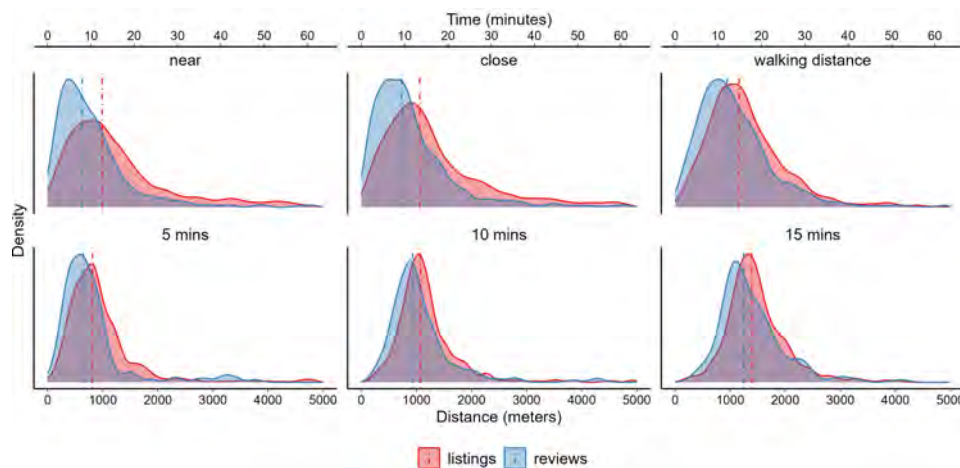
calculate the shortest walking distance between each Airbnb and entity coordinates along OpenStreetMap’s pedestrian network. We use a maximum threshold value of 5,000 meters to remove any outliers in the data. We do not expect individuals to walk distances exceeding 5,000 meters since the largest STE used in our analysis is 15 minutes.

For RQ2, we generate density plots to examine how walking distances are distributed across STE groups and tags for listings and reviews. We use a secondary axis to display walk time, calculated using an average walking speed of 1.31 meters per second [14]. Differences in these distributions are assessed using a pairwise Mann-Whitney U test, a non-parametric statistical test commonly used for data that is not normally distributed [12]. We use this test to determine whether the differences in distributions are statistically significant (RQ3). Finally, for RQ4 we plot the average differences across the widely used NYC neighborhoods dataset by the non-profit BetaNYC.⁵

3 Results

Looking at Figures 2 and 3, we see that qualitative distance measures like “nearby” or “walking distance” can indeed be quantified using the methods detailed above (RQ1). By comparing these quantifications across listings and reviews, we discover that the former tend to exaggerate proximity more than the latter (RQ3). However, across both types of data, claimed walking times (5, 10, and 15 mins) were distributed widely across the actual walking times (RQ2). Walking distances from listings were on average 12 minutes when the stated distance was 5 minutes, 15 minutes for 10 minutes, and 18 minutes for 15 minutes. Walking distances from reviews, by contrast, were closer to the actual claim: 8 minutes for 5 minutes, 12 minutes for 10 minutes, and 15 minutes for 15 minutes. These differences are also reflected in how words like “near,” “close,” and “walking distance” are deployed on average: For listings, these are close to 15 minutes of walking, while only 10 for reviews. Furthermore, turning our attention to figure 3, words like “nearby” are always closer for parks than for schools and tourist attractions. Again, listings consistently exaggerate proximity across these three toponymic categories but the general pattern also holds: “nearby” parks are only 10 minutes away for reviews and 13 minutes away for listings, whereas schools are 12 and 15 minutes away and tourist attractions 15 and 17 minutes.

⁵ <https://data.beta.nyc/dataset/peidiacities-nyc-neighborhoods/resource/35dd04fb-81b3-479b-a074-a27a37888ce7>



■ **Figure 3** Walking distances and times for six different spatio-temporal entities in the data. Again, the means are consistently lower for the reviews than the listings. The differences are particularly pronounced for the vague STE qualifiers (top row).

We also found that there were statistical differences across neighborhoods (RQ4). Looking at Figure 1b, listings in Manhattan tend to exaggerate consistently compared to reviews: here, it seems, everything is “nearby.” This trend is reversed only in Lower Manhattan. Outside of Manhattan, the visually distinct spatial clusters are more mixed. As we move further out of the city center, the differences become more extreme in both directions. While this is interesting to note, issues with data sparsity and outliers might be part of the explanation. Nonetheless, these patterns require further investigation. Are listings in less attractive neighborhoods more prone to exaggerate when they talk about distance? How might these patterns correlate with the cultural and economic hierarchy between different areas [11]?

4 Discussion and conclusions

In this paper, we demonstrated how spatial entities and relations can be extracted from textual descriptions of reviews and listings from Airbnb (RQ1). Claimed walking distances do not reflect real world walk-times (RQ2), but the exaggeration is more extreme in listings than reviews (RQ3). While these differences seem to be spatially clustered (RQ4), the exact nature of these clusters remains to be investigated. Although these results are preliminary, they offer a first step towards exploring the dynamics between the representation of spatial relations and place-making.

There are notable limitations to our approach. First, it remains to be seen whether our trained models would generalize well to other settings. Second, our model for extracting spatial entities and our method for parsing spatial relations are still imperfect, introducing a margin of error in the results. Third, there are sparsity issues with some of our annotated data, which is reflected in the uneven F1-scores between labels (see Table 1).

These reservations notwithstanding, we have shown how to quantify and extract vague spatial relations from text data. Moreover, we have demonstrated that there are consistent and statistically significant differences between listings and reviews – that is, between hosts and guests – in their representations of spatio-temporal relations. In this way, the results presented here open up a new vantage point to studying representations of spatial relations through geocoded text data. For example, by exploring how changes in these representations

change over time, they could be related to indices of gentrification. Furthermore, these methods could be expanded beyond the scope of Airbnb data to analyze representations of space in a number of textual contexts: short- versus long-term real estate descriptions, other forms of tourism literature, and even fictional literature.

References

- 1 Kyle Barron, Edward Kung, and Davide Proserpio. The Effect of Home-Sharing on House Prices and Rents: Evidence from Airbnb. *Marketing Science*, 40(1):23–47, October 2020.
- 2 Pierre Bourdieu. *Distinction: A Social Critique of the Judgement of Taste*. Harvard University Press, Cambridge, MA, 1984.
- 3 Lucie Cadorel, Denis Overal, and Andrea G. B. Tettamanzi. Fuzzy representation of vague spatial descriptions in real estate advertisements. In *Proceedings of the 6th ACM SIGSPATIAL International Workshop on Location-based Recommendations, Geosocial Networks and Geoadvertising*, pages 1–4, Seattle Washington, November 2022. ACM. doi:10.1145/3557992.3565994.
- 4 Robbin Deboosere, Danielle Jane Kerrigan, David Wachsmuth, and Ahmed El-Geneidy. Location, location and professionalization: a multilevel hedonic analysis of Airbnb listing prices and revenue. *Regional Studies, Regional Science*, 6(1):143–156, January 2019.
- 5 Curdin Derungs and Ross S. Purves. Mining nearness relations from an n-grams Web corpus in geographical space. *Spatial Cognition & Computation*, 16(4):301–322, October 2016.
- 6 E. Goffman. *The Presentation of Self in Everyday Life*. Anchor Books, New York, NY, 1959.
- 7 Zhiheng Huang, Wei Xu, and Kai Yu. Bidirectional LSTM-CRF Models for Sequence Tagging, August 2015. arXiv:1508.01991 [cs]. doi:10.48550/arXiv.1508.01991.
- 8 F. Jameson. *Postmodernism, Or, The Cultural Logic of Late Capitalism*. Duke University Press, 1991.
- 9 JD. Lafferty, A. McCallum, and Fernando C. N. Pereira. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the Eighteenth International Conference on Machine Learning, ICML '01*, pages 282–289, San Francisco, CA, USA, June 2001. Morgan Kaufmann Publishers Inc. doi:10.5555/645530.655813.
- 10 Kevin Lynch. *The Image of the City*. MIT Press, Cambridge, MA, 1964.
- 11 David Madden. Neighborhood as Spatial Project: Making the Urban Order on the Downtown Brooklyn Waterfront. *International Journal of Urban and Regional Research*, 38(2):471–497, 2014. _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/1468-2427.12068>.
- 12 H. B. Mann and D. R. Whitney. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics*, 18(1):50–60, March 1947. Publisher: Institute of Mathematical Statistics. doi:10.1214/aoms/1177730491.
- 13 G. McKenzie and Y. Hu. The “Nearby” exaggeration in real estate. In *Proceedings of the Cognitive Scales of Spatial Information Workshop, L’Aquila, Italy*, pages 4–8, 2017.
- 14 Elaine M. Murtagh, Jacqueline L. Mair, Elroy Aguiar, Catrine Tudor-Locke, and Marie H. Murphy. Outdoor Walking Speeds of Apparently Healthy Adults: A Systematic Review and Meta-analysis. *Sports Medicine*, 51(1):125–141, January 2021.
- 15 L. A. Ramshaw and M. P. Marcus. Text Chunking Using Transformation-Based Learning. In Susan Armstrong, Kenneth Church, Pierre Isabelle, Sandra Manzi, Evelyne Tzoukermann, and David Yarowsky, editors, *Natural Language Processing Using Very Large Corpora*, Text, Speech and Language Technology, pages 157–176. Springer Netherlands, Dordrecht, 1999.
- 16 Alexandria Ravenelle. A return to Gemeinschaft: Digital impression management and the sharing economy. In J. Daniels, K. Gregory, and TM Cottom, editors, *Digital sociologies*, pages 27–45. Bristol University Press, 1 edition, November 2016. URL: 10.2307/j.ctt1t89cfr.
- 17 Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter, February 2020. arXiv:1910.01108 [cs]. doi:10.48550/arXiv.1910.01108.

- 18 David Wachsmuth and Alexander Weisler. Airbnb and the rent gap: Gentrification through the sharing economy. *Environment and Planning A: Economy and Space*, 50(6):1147–1170, September 2018. doi:10.1177/0308518X18778038.

A Appendix

■ **Table 1** Summary of NER label frequencies in the training data, in the overall data, and performance metrics (F1, Recall, and Precision) for the DistilBERT-CRF model. The plain DistilBERT model produced similar numbers.

		N	N	N			
	Label	Annotated	Predicted	Predicted	F1	Rec.	Prec.
		(all)	(listings)	(reviews)			
1	TN:NEIGHBORHOOD	2265	66211	39914	0.872	0.877	0.867
2	TN:BOROUGH	1677	30014	40611	0.932	0.944	0.920
3	TN:CITY	1000	20097	46674	0.941	0.955	0.928
4	TN:STREET	552	21259	7409	0.681	0.675	0.687
5	TN:STATION	543	19828	8233	0.582	0.621	0.547
6	TN:TOURIST_ATTR	615	21619	6619	0.619	0.646	0.593
7	TN:PARK	532	19201	9548	0.893	0.941	0.850
8	TN:SCHOOL	127	3132	943	0.516	0.457	0.592
9	TN:BUSINESS	730	24059	9623	0.718	0.742	0.695
10	TN:OTHER	347	6092	3029	0.413	0.415	0.411
11	SPAT_TEMP_ENT	6643	197089	203545	0.690	0.708	0.672
12	TRANSIT	4168	126360	105646	0.787	0.806	0.768
13	GEOG_ENTITY	6663	184825	261947	0.806	0.812	0.800
14	HOST_BUILDING	915	29364	12391	0.426	0.442	0.411
	Overall	26777	769150	756132	0.756	0.771	0.742

A.1 Sampling

To sample the training data, we used the following stratified disproportionate sampling strategy:

1. Per neighborhood, all listings are included if there are 5 or fewer.
2. In neighborhoods with more listings than that, the sample for the neighborhood is 5 listings + 0.5%.
3. Each listing has 1 review sampled, but many listings have no reviews.

Sampling like this, we could ensure that all neighborhoods were represented in the training data. However, for the review data that the trained model extracted NER labels from, we used no spatial stratification, which is potentially reflected in the results. Future work should use the entire dataset of reviews or take a spatially stratified sample.

A.2 Models

We trained three different models on the annotated data: 1) DistilBERT [17] with a linear classification layer, 2) DistilBERT with a conditional random fields (CRF) [9] layer prior to the linear classifier, and 3) DistilBERT with a CRF and BiLSTM layer prior to the linear classifier [7]. For all these models, we used a 10/90 test-train split. Between the models, the

The listing is inaccurate about the **location** `GEOG_ENTITY`, the **distance to** `SPAT_TEMP_ENT` **Manhattan** `TN:BOROUGH` is **at least 70 minutes by public transport** `SPAT_TEMP_ENT` and **45 minutes by car minimum** `SPAT_TEMP_ENT`, it's **an hour walk to** `SPAT_TEMP_ENT` the **nearest** `SPAT_TEMP_ENT` **subway station** `TRANSIT`. But overall a lovely **place** `HOST_BUILDING` and a nice **neighborhood** `GEOG_ENTITY`

■ **Figure 4** The annotation interface of Prodigy. This annotated review references several different types of entities related to place and spatial relations.

DistilBERT model with a CRF layer but without the BiLSTM layer performed the best, with an overall F1-score of 0.756. Almost similar results were achieved with the DistilBERT model, with a 0.752 F1-score. To keep results reproducible, all downstream tasks were performed with this model. While these F1-scores might seem on the low side, it was much higher for many of the classes in the data, as can be seen in table 1. The final models for all three architectures were trained over five epochs with a 1×10^{-4} learning rate, 1×10^{-5} weight decay, gradient clipping, and early stopping. All models were implemented in PyTorch⁶ using pretrained DistilBERT models from HuggingFace⁷ and using additional IOB-chunking [15].

For an example of the annotation interface and, consequently, the data that was given to the models, see figure 4.

A.3 Relationship extraction

To extract the dependencies between Spatio-Temporal Entities (STEs) and toponyms, we proceed in the following way: For each document in our corpus, we extract dependencies using the spaCy Python library⁸, with entities recognized as toponyms merged into single tokens. We next identify all the dependents for all tokens for each document, using these relations to build a directed graph of each document. Given this graph, we filter for nodes that are labeled STE and remove any edges that point to this node. Next, we find the weakly connected subgraphs that remain after removing these edges, giving us a set of graphs with at most one STE node each and n nodes with other labels, including toponyms. Now, each of these other nodes is a dependent of an STE node and we can pair each toponym-labeled node with the STE of the subgraph.

⁶ <https://pytorch.org/>

⁷ <https://huggingface.co/>

⁸ <https://spacy.io/>

Smarter Than Your Average Model - Bayesian Model Averaging as a Spatial Analysis Tool

Chris Brunsdon ✉ 

National Centre for Geocomputation, Maynooth University, Ireland

Paul Harris ✉

Rothamsted Research, Harpenden, UK

Alexis Comber ✉

School of Geography, University of Leeds, UK

Abstract

Bayesian modelling averaging (BMA) allows the results of analysing competing data models to be combined, and the relative plausibility of the models to be assessed. Here, the potential to apply this approach to spatial statistical models is considered, using an example of spatially varying coefficient modelling applied to data from the 2016 UK referendum on leaving the EU.

2012 ACM Subject Classification Mathematics of computing → Bayesian nonparametric models

Keywords and phrases Bayesian, Varying coefficient regression, Spatial statistics

Digital Object Identifier 10.4230/LIPIcs.GIScience.2023.17

Category Short Paper

Supplementary Material *Text:* https://drive.google.com/file/d/1HHb5uEsGX0Qu61hIwWd-8BgjTrGzn6R-/view?usp=share_link

1 Overview

Imagine that you are waiting for a taxi and it is already slightly late. You are concerned that you will miss a train, and want to estimate how long you will need to wait. A number of scenarios could cause the delay. For example: The taxi is stuck in traffic; There was an administrative error and the booking service gave the taxi driver the wrong time; The taxi was involved in a road accident; and so on. In each case a number of factors effect the expected delay - but the factors are not the same in each scenario. However your main concern is the delay time, regardless of the scenario. This is a similar problem to those which Bayesian Model Averaging (BMA) may be used to address.

If you had models encompassing k scenarios based on past data D - say $\{M_1 \cdots M_k\}$ intended to predict the delay time T , and posterior beliefs in each scenario being correct: $\{\Pr(M_1|D) \cdots \Pr(M_k|D)\}$ you could obtain the predictive distribution of T given D as a weighted average of the individual predictive distributions obtained from each model as

$$\Pr(T|D) = \sum_{i=1,k} \Pr(T|M_k, D)\Pr(M_k|D).$$

This in essence is Bayesian Model Averaging (BMA) – if we have a number of competing models with at least one quantity of interest that all have in common, and relative likelihoods of each of them being the correct model, we can obtain a posterior distribution of the quantity of interest by averaging them using the likelihoods as weights.

Up to this point, there is nothing exclusively spatial about this process, but it can be a powerful tool for assessing and utilising spatial models. For example, the competing models could be:



© Chris Brunsdon, Paul Harris, and Alexis Comber;
licensed under Creative Commons License CC-BY 4.0

12th International Conference on Geographic Information Science (GIScience 2023).

Editors: Roger Beecham, Jed A. Long, Dianna Smith, Qunshan Zhao, and Sarah Wise; Article No. 17; pp. 17:1–17:6

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1. Spatial regression models using different spatial weight matrices.
2. Spatially Varying coefficient regression models where different parameters have fixed or spatially varying coefficients in each model.
3. Spatial trend models with differing map projections (eg. a cartogram vs. national grid coordinates)

In general, this approach can be used for any parameter that is common to all models, or a predicted dependent variable – so if one were interested a particular regression coefficient, its posterior distribution could be considered in terms of various models containing this coefficient. A key advantage of this approach is that while many other approaches (eg stepwise regression, best AIC, best cross validation score) have a workflow to select a single “best” model, this averages over all possibilities on the basis of relative evidence. In particular when several models all perform similarly well, this approach makes use of information from all of them, rather than discarding all but one.

2 A Brief Description of Computational Methodology

The approach to computing $\Pr(M_i|D)$ – a crucial stage in BMA - is to firstly compute $\Pr(D|M_i)$ – then, via Bayes’ Theorem, we have

$$\Pr(M_i|D) = \frac{\Pr(D|M_i)\Pr(M_i)}{\sum_j \Pr(D|M_j)\Pr(M_j)}.$$

Each model M_i will have its own parameter vector Θ_i - although the respective Θ_i may differ in length and form between models. Standard statistical models typically specify $\Pr(D|\Theta_i, M_i)$ - but here we are interested in the *marginal* probability of the observed data D across all possible Θ_i values for each M_i , weighted by their prior probabilities. That is

$$\Pr(D|M_i) = \int_{\Theta_i} \Pr(D|\Theta_i, M_i)\Pr(\Theta_i|M_i) d\Theta_i.$$

This is sometimes referred to as the marginal posterior probability of D given M_i . Although the right hand side expression cannot usually be derived analytically, two broad approaches may be taken:

1. Approximation.
2. Monte Carlo Simulation.

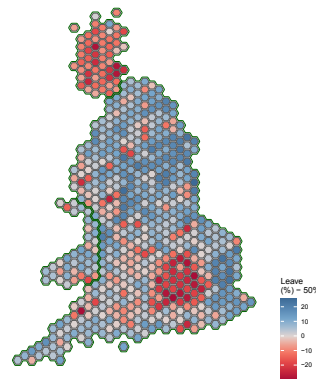
Approximation is generally quicker and less “resource hungry” to evaluate, but less accurate. A usual strategy for approximation is based on the Bayesian Information Criterion (BIC) [4] for model M_i . If $\hat{\Theta}_i$ is the maximum likelihood estimate for Θ_i for M_i , and \hat{L} is the value of the likelihood corresponding to $\hat{\Theta}_i$, n is the sample size, and k is the dimension of Θ_i then

$$\text{BIC} = k \log(n) - 2 \log(\hat{L})$$

and for larger n it can be shown that

$$\Pr(D|M_i) \approx \exp\left(-\frac{\text{BIC}}{2}\right).$$

Finally, for the parameter(s) of interest, say $\theta_i \subset \Theta_i$ for a given model M_i the posterior distribution can be approximated via Laplace’s approximation [1]. The posterior distribution for θ_i may be approximated as having a multivariate normal distribution with a variance-covariance matrix equal to the Hessian of the posterior likelihood function, with the maximum



■ **Figure 1** Leave Vote (%) by Parliamentary Constituency.

■ **Table 1** Variables Used in Referendum Outcome Modelling.

Variable	Description
Leave	Percentage of “Leave” votes for each constituency (Dependent Variable).
Born_uk	Percentage of electorate born in the UK.
Age_65_plus	Percentage of electorate aged 65 or older.
Turnout	Percentage of electorate who voted in the 2015 general election.
Christian	Percentage of electorate stating their religion as “Christian”.

likelihood estimators of θ_i as mean values. For a scalar θ_i this suggests that the marginal posterior distribution may be estimated as a normal distribution with the maximum likelihood estimate $\hat{\theta}_i$ as its posterior mean, and $SE(\hat{\theta}_i)$ as its posterior standard deviation. The BMA may then be approximated as a mixture of the k Normal distributions with $\Pr(M_i|D)$ as the weight for M_i .

In this study, the example will use the BIC-based approach, and so attention will be focused on this method.

3 Example: The UK’s 2016 Referendum on Leaving the EU

On June 23rd 2016, the United Kingdom held a referendum regarding its then membership of the European Union. Voters were offered two choices: “Leave the European Union” (Leave) or “Remain a member of the European Union” (Remain). The outcome was a 51.9% majority in favour of “Leave”, although a hexagonal cartogram map of voting by Parliamentary Constituencies in England, Scotland and Wales (Figure 1) suggests this overall figure conceals notable regional patterns. This leads to a further question: if the voting patterns themselves show strong regional patterns, do the *drivers* of these outcomes also vary geographically?

To investigate this, a number of variables were obtained (from the `parlitoools` R package) [3], recorded at the Parliamentary Constituency geographical unit – listed in Table 1. The UK census-based variables (`born_uk`, `age 65+`, and `Christian`) were recorded in the 2011 UK Census – this being the latest Census held in the UK prior to the referendum.

17:4 Bayesian Model Averaging for Spatial Analysis

The key questions for each variable are whether they influence the leave vote; and if so then does the direction and magnitude of this influence vary geographically? To investigate this, for each variable it is possible to include it in a model with a fixed linear coefficient $\beta \times \text{Variable}$ or a geographically varying coefficient $\beta(u, v) \times \text{Variable}$ where (u, v) is the centroid of each parliamentary constituency, or to omit it from the model.

To investigate this, the R package `mgcv` was used to fit every permutation of these kinds of model. For each of the four predictor variables there were three possibilities - omit the variable from a regression model, include with a fixed linear coefficient, or include with a spatially varying coefficient. In the latter case a thin-plate spline approach was used (although other options could be chosen). In the R formula notation, an example of a model might be

```
Leave ~ s(u,v,by=Born_uk,bs='tp') + Turnout
```

suggesting a model where the coefficient for `Turnout` was fixed, that for `Born_uk` varied, and the other variables were omitted. This yields $3^4 = 81$ models. In addition to this, each model was fitted with both fixed and varying intercept terms, and with the coordinates (u, v) based on location on the cartogram and physical (UK National Grid) location. Thus there are 4 variants on each model, resulting in $81 \times 4 = 324$ possible models altogether. In the marginal likelihood approach, there is no requirement that models be nested, so all 324 models can be considered. Here the `mgcv` package offers Bayesian Information Criterion methods (BIC) for `gam` model objects, and so the BIC based approximation will be used here. Using this approach, all models with a posterior probability ≥ 0.01 are listed in Table 2.

The most likely model includes all variables with the intercept and the `Born_uk` coefficient being modelled as thin plate splines, and the remaining variables having fixed linear coefficients. The geographical coordinates for the splines are based on the cartogram, rather than physical space. However, reading the $\Pr(M|D)$ column in the table suggests that this model is the correct one is a little under two thirds. The possibility of one of the “runners up” being correct is non-trivial. In the next model in the table (probability around one in five) `Born_uk` has a fixed coefficient - but also although the intercept term is varying, the coordinates are now based in *physical* space.

The `Intercept` term has a spatially varying coefficient in all of the three most probable models. These three models dominate the posterior marginal probabilities, totalling around 0.95 of all possibilities. These surfaces are shown in Figure 2. The Bayesian model average surface (over all possible models) for intercept is given by

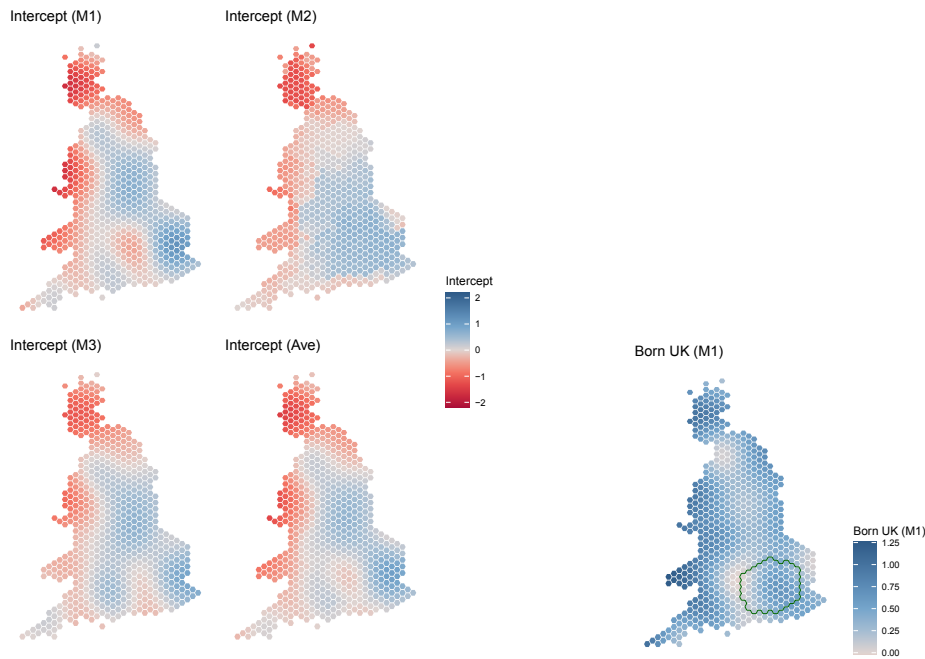
$$\beta_{0*}(u, v) = \sum_{i=1 \dots 324} \Pr(M_i|D) \beta_{0i}(u, v)$$

where $\beta_{0i}(u, v)$ is the intercept coefficient for model i . For models where the intercept is constant, $\beta_{0i}(u, v)$ is a constant w.r.t. (u, v) . This is shown as the fourth map in Figure 2 on the LHS map quartet.

In these models all variables - as listed in Table 1 - are standardised to have mean zero and standard deviation 1 prior to analysis. For the intercept term, this gives the standardised value for the `Leave` variable assuming all other predictors are at their mean value. It is not a direct measure of overall tendency to vote “Leave” or otherwise - more of a measure of geographical effects not accounted for by current variables in the model. On this basis, there seems to be among other things a “Scotland effect” and a “West London effect” (although this is not apparent in the second most probable model, which uses physical coordinates rather than cartogram). Once the models are averaged the West London effect remains, although muted.

■ **Table 2** Models with Highest Posterior Probabilities.

Intercept	Born_uk	Age 65+	Turnout	Christian	Coords	Pr(M D)
Spline	Spline	Fixed	Fixed	Fixed	Cartogram	0.637
Spline	Fixed	Fixed	Fixed	–	Physical	0.205
Spline	Fixed	Fixed	Fixed	Fixed	Cartogram	0.109
Spline	Fixed	Fixed	Fixed	Fixed	Physical	0.045



■ **Figure 2** The intercept and born_uk terms by parliamentary constituency (Great Britain).

The coefficient for `born_uk` can also be mapped. This is shown in Figure 2 (RHS). The values are calculated using the formulae above. Of note here is perhaps that in a region to the west of London, `born_uk` seems to have little influence on the outcome than in much of the country where higher values suggest a **Leave** majority is more likely.

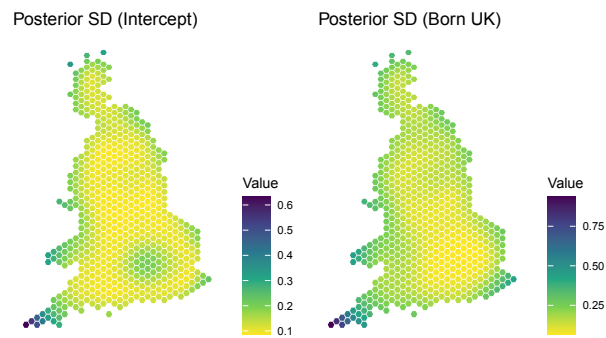
It is also possible to map the marginal posterior standard deviation for these parameters, after model averaging. These are computed using the formula

$$[\text{PSD}(\beta_*(u, v))]^2 = \sum_{i=1 \dots 324} \text{Pr}(M_i|D) [\text{PSD}(\beta_i(u, v))]^2$$

and are shown in Figure 3. Notable in both cases is the “edge effect” where the PSD is high near to the coastal areas. Also of note is the raised PSD in the London area.

4 Discussion

The BMA approach provides a number of useful tools. It provides a means of assessing the viability of competing models, by providing posterior probabilities of each being the correct model. This can be thought of as similar to hypothesis testing, but it treats hypotheses symmetrically, and can handle more than two competing hypothesis. It also provides means



■ **Figure 3** Posterior Standard Errors for `Intercept` and `born_uk`.

of combining competing models to investigate parameters common to all models, in the presence of uncertainty as to which model is correct. The example here used an approximate approach that is convenient, as it can be achieved using standard R tools. More accurate approaches are also possible via techniques such as *Bridge Sampling* – see [2] for example.

References

- 1 Robert E. Kass, Luke Tierney, and Joseph B. Kadane. Laplace’s method in Bayesian analysis. *Contemporary Mathematics*, 115:89–100, 1991. doi:10.1090/conm/115/07.
- 2 X. Meng and S. Schilling. Warp bridge sampling. *Journal of Computational and Graphical Statistics*, 11(3):552–586, 2002. doi:10.1198/106186002457.
- 3 Evan Odell. *parlitoools: Tools for analysing UK politics in R*, 2017. R package version 0.4.1. doi:10.5281/zenodo.591586.
- 4 Gideon Schwarz. Estimating the Dimension of a Model. *Annals of Statistics*, 6(2):461–464, 1978. doi:10.1214/aos/1176344136.

Anonymous Routing Using Minimum Capacity Clustering

Maike Buchin ✉

Ruhr University Bochum, Germany

Lukas Plätz ✉

Ruhr University Bochum, Germany

Abstract

We present a framework which allows one to use an online routing service and get live updates without revealing the sensitive starting and ending points of one’s route. For that, we obfuscate the starting and ending locations in minimum capacity clusters and reveal only the route between these clusters. We compare different anonymous clustering strategies on positions in the network with efficient approximations and analyse the impact of the anonymisation on the route. We experimentally evaluate the effect of the anonymisation scheme in real-world settings.

2012 ACM Subject Classification Security and privacy → Pseudonymity, anonymity and untraceability

Keywords and phrases Anonymity, approximation Algorithms, directed Networks, minimum capacity Clustering, Privacy

Digital Object Identifier 10.4230/LIPIcs.GIScience.2023.18

Category Short Paper

Supplementary Material *Software (Code and Data):*

<https://gitlab.ruhr-uni-bochum.de/plaetlsv/giscience23/-/tree/FrechetAbstand>

Funding *Lukas Plätz*: The work was supported by the PhD School “SecHuman – Security for Humans in Cyberspace” by the federal state of NRW.

1 Introduction

Services often utilise personal routing data to offer traffic information, but it can be achieved using anonymised data. We can protect the sensitive part of our data by trading in a small amount of convenience. Anonymising the routing data enables us to use it in scientific research and redistribute it. The central idea is that the two endpoints of a route determine the shortest path. This means that the shortest paths only is helpful in re-identifying the starting and ending locations. By obfuscating these locations, we can protect privacy while sharing the remainder of the route for the public and personal benefit.

The concept of “ k -anonymity” was first introduced by Sweeney [6]. It guarantees that each subject cannot be distinguished from less than $k - 1$ other subjects. So finding a good k -anonymisation can be viewed as a clustering problem, with clusters requiring a minimum capacity of k . We k -anonymise locations in the network by clustering them.

To achieve k -anonymity, we adopt a concept from the routing literature introduced by Bast et al. [3]. In long-distance travel, routes around a starting location pass through a small set of nodes near the start. These nodes, known as *transit nodes*, reduce the search space and speed up the shortest path computation. We use a variation of this concept to anonymise the routing of a person. By computing transition nodes for each cluster (possibly depending on the cluster to be routed to) and routing through them, we can ensure that the path between these nodes remains the same for all starting and ending locations within the cluster.



© Maike Buchin and Lukas Plätz;
licensed under Creative Commons License CC-BY 4.0

12th International Conference on Geographic Information Science (GIScience 2023).

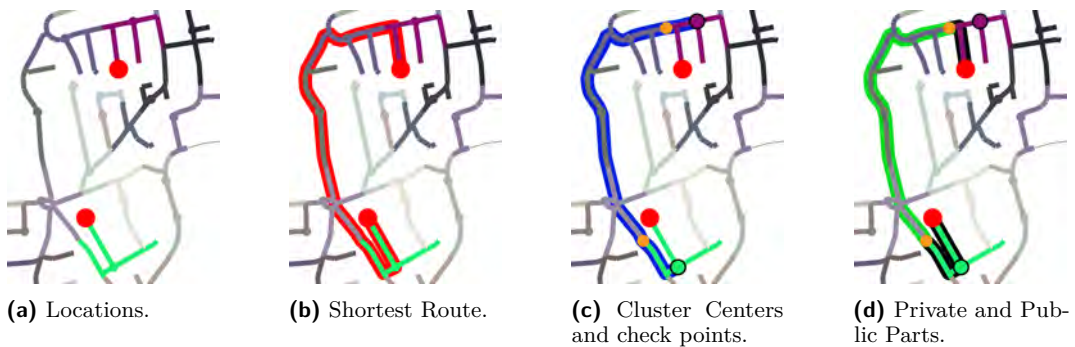
Editors: Roger Beecham, Jed A. Long, Dianna Smith, Qunshan Zhao, and Sarah Wise; Article No. 18; pp. 18:1–18:6

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Given clusters of at least k locations, one could find the transit nodes between two clusters. Utilising these transit nodes per cluster would keep the travel time the same. However, it would weaken anonymity, as a (specific) transit node (of several transit nodes of a cluster) may reveal in which part of the cluster the point lies. Therefore, we instead decided to actively lead the route through *check points* on the boundary of the clusters. This re-routing introduces some additional travel time, denoted as Δ . We will show that the maximum of Δ can be upper bounded with the radius of the clusters. Moreover, for a wide range of k , the mean value of Δ is insignificant in daily use. Additionally, since the check points are on the boundary of the clusters, most of the routes can be shared. See Figure 1 for an example. Routes within a cluster will not be anonymised within our framework as they are too short for gains through online services and would not use any check points.



■ **Figure 1** shows an example of the anonymisation strategy. In the road network, streets from the same cluster have the same colour. In red are the locations and the shortest route. Next, we look at the centres shown as dots in colour for their cluster. With the shortest route in blue between them, get the check points in orange on the boundary of the clusters. Lastly, we compute the private in the black and the public in the green part of the anonymous route between the red locations.

We discuss four clustering strategies that differ in their setting and optimisation criterion while achieving a k -anonymous clustering. Later we will compare them on their impact on Δ . First is the *r-gather clustering* problem, which Aggarwal et al. [1] introduced. Here, the objective is to find clusters where each cluster contains a minimum of r points. The cluster's centre determines its radius, and the goal is to minimise the maximum radius across all clusters. They showed that this problem is NP-hard and gave a polynomial time algorithm to compute a 2-approximation, i.e. the radius is at most two times the optimal radius.

Armon [2] presents two variants of *r-gathering* that interest us. In the *r-gathering setting* – in contrast to *r-gather clustering* – the centres are chosen from a different set than the points to be clustered. The first one, called *min-max r-gathering*, minimises the maximum radius of the clusters. It is an NP-hard problem, and they presented a 3-approximation in the maximum radius in $O(n(m + r + \log n))$ time. The second strategy, *min-sum r-gathering*, minimises the sum over the distances to the centre. They showed that the problem is NP-hard and gave a $2r$ approximation in $O(n(m + r + \log n))$ time. With *r-gather* and *min-max r-gathering*, we can compute bounds to time lost by our anonymisation. However, *min-sum r-gathering* should lead to a better mean Δ than the other strategies.

Hauert et al. [5] introduced the *k-Anonymous Steiner Forest* for the problem of location clustering. Here they compute the optimal clustering where a cluster has to pay to the length of the street connecting them. They gave an algorithm with an approximation factor of 2 and a runtime of $O(nm)$. They applied their strategy to clustering places in a street network. However, their optimisation criterion does not align with ours, as the cost of the edge is only paid once. However, we will compare our location clustering with their result.

Brauer et al. recently presented a solution to a similar problem [4] which builds on the clustering strategy of Haunert et al. [5] to truncate trajectories. However, they only considered geometric clues¹. With that, the trajectories leak information about the start and end points. The attacker model is heavily constrained because it cannot use the knowledge of the existing network. However, their strategy can be used to anonymise existing databases of trajectories, and they do not need the shortest paths.

We present a framework for k -anonymous online routing. We argue that under our assumptions (that all users use the shortest route), retrieving the start or ending from the route is impossible, even if the attacker knows the model of obfuscation, the clusters and the network. We bound the impact on the travel time by this framework and present a polynomial time algorithm to minimise the impacted travel time. We demonstrate the practicality of our framework with experimental results for German cities.

2 Anonymization Scheme

We will use the road network, given as a directed embedded graph $G = (V, E)$ and information on travel time $t(e)$ and population distribution $p(e)$ over the edges of the network first to compute a clustering on the edges. We then anonymise the shortest path between two points by calculating the shortest path between the cluster centres that encompass these points. We exclude the portion of the path within the clusters and replace it with the shortest route from the starting point to the path and from the path to the destination.

This setting brings two challenges with it. First, typically the vertices of a graph are clustered, whereas clustering of edges is rare. To use the point clustering techniques, we have to adapt our graph. For that, we use the directed line graph, which has a node for each edge in the directed graph and maintains the connectivity by introducing a directed edge for each pair of edges in the directed graph if the first edge ends at the start of the second edge.

Secondly, road networks are directed graphs which do not come with a canonical metric. We decided to use the length of the minimal cycle of a list of items as our distance function. We use this because it gives a symmetrical distance measure for a list of length two and satisfies the triangle inequality. Also, roundtrips are meaningful in our settings. We used items as a stand-in for vertices and edges. To make this distance measure a metric, we define cycle $[p, p]$ to have length 0. As we will primarily discuss minimal cycles, we use the list notation $[a, b, \dots]$ if we mean the shortest cycle using these objects in that order. The distance function d gives us the length of the minimal cycle.

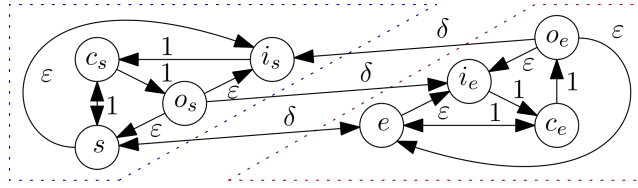
For a cluster C in the clustering \mathcal{C} , we denote its centre as c_C . For two clusters C, C' , we get the check points from the minimal cycle $[c_C, c_{C'}]$. The entry check point is the first node $i_C \in C$ on the minimal cycle, coming from $c_{C'}$. The exit check point is the last node $o_C \in C$ on the minimal cycle, coming from c_C . So for the shortest cycle $[s, e]$ with $s \in S$, $e \in E$ and $S, E \in \mathcal{C}$, this leads to the anonymized cycle $\mathcal{A}(s, e) := [i_s, s, o_s, i_e, e, o_e]$.

We define the radius $R(C)$ of a cluster C as $\max_{p \in C} d([c_C, p])$. Furthermore, $\mathcal{R} := \max_{C \in \mathcal{C}} R(C)$ denotes the maximal radius of the clusters in a clustering.

Now we can define the function $\Delta(s, e)$ from the introduction as $d(\mathcal{A}(s, e)) - d([s, e])$.

► **Lemma 1.** *Given the maximum radius \mathcal{R} of a clustering, $4\mathcal{R}$ bounds the maximum extra time Δ introduced by the anonymisation scheme \mathcal{A} .*

¹ i.e. the closest location and a wedge in the last direction of the trajectory



■ **Figure 2** Example showing the tightness of the upper bound as stated in the lemma 1. If we pick c_s and c_e as the centres, we get a 4-gather with radius $2 + \epsilon$. The minimal cycle distance between s and e is 2δ . The anonymised cycle distance is $8 + 2\delta$. This gives us the tightness for ϵ approaching 0.

Proof. By definition, we have $\Delta := d(\mathcal{A}(s, e)) - d([s, e])$. When we insert the centres c_s and c_e into the anonymised cycle, we only make it longer but also can drop the check points as they lie on the shortest cycle between c_e and c_s .

$$\Delta \leq d([i_s, c_s, s, c_s, o_s, i_e, c_e, e, c_e, o_e]) - d([s, e]) = d([c_s, s, c_s, c_e, e, c_e]) - d([s, e])$$

If we now insert s and e between c_s and c_e we can split the long cycle into smaller ones,

$$d([s, c_s, s, c_s, s, e, c_e, e, c_e, e]) = d([c_s, s]) + d([c_s, s]) + d([s, e]) + d([c_e, e]) + d([c_e, e]).$$

The length of the smaller cycles within a cluster is bounded by \mathcal{R} . Thus, we get $\Delta \leq 4\mathcal{R}$. ◀

Remarkably, the upper bound $4\mathcal{R}$ is tight. Figure 2 shows an example for that. Also, Δ can be bounded by the actual radius of the starting and ending clusters.

We conclude that minimising the maximum radius of the clustering is a suitable proxy/substitute for anonymising with a small impact on travel time.

3 Experimental Results

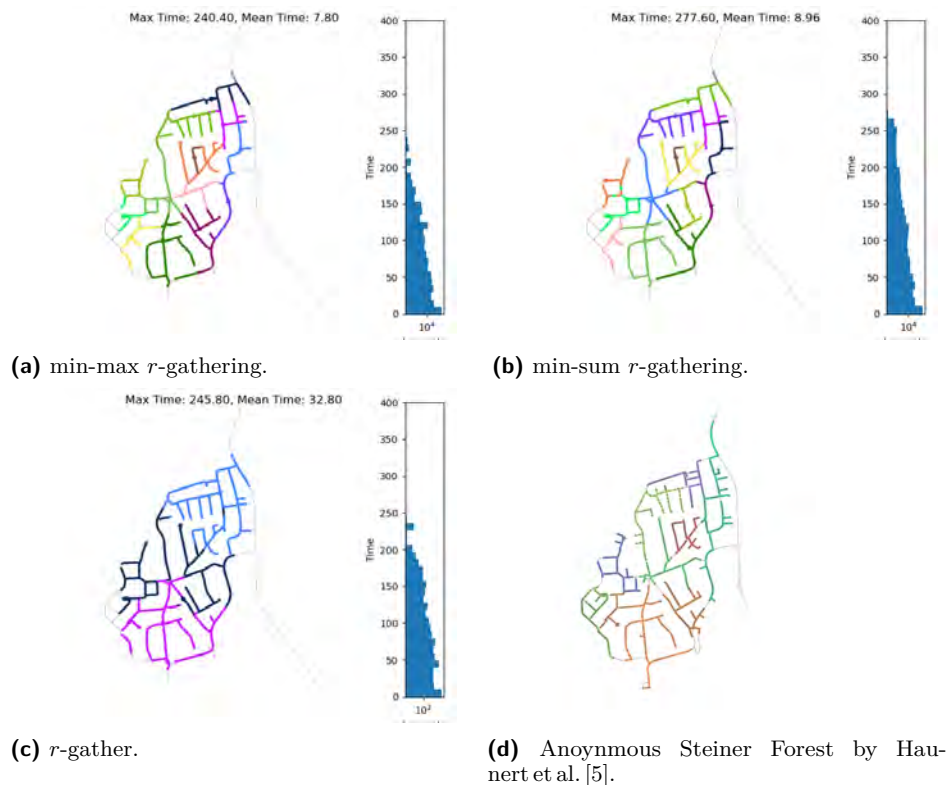
We tested our anonymisation scheme in several cities in Germany. We used the data from OpenStreetMap for the network and the German census data to estimate the number of people living next to the streets. We import the street network with the travel time for the edges from OpenStreetMaps. The German census [7] from 2011 provided a 100 m times 100 m square grid of people living in each cell.² We distributed each square’s population evenly on every curve in that square for a realistic distribution.³

To use the r -gather clustering, we used a line digraph of the network to switch the roles of edges and vertices. Because all clustering strategies assume that each point has equal weight, we used a multigraph with as many edges for a street as people.

Our primary focus was to analyse Δ for the different strategies and r , as this bounds the detour induced by the anonymisation scheme. For that, we computed the shortest and anonymised path for every pair of vertices. With that, we calculated the values of Δ . We empirically found that the maximum Δ is often close to two times the maximum radius of the clusters. This could be explained by the fact that most edges of a street network are undirected, and in that case, two times the maximum radius is the upper bound. Nevertheless, there are instances where it gets close to the upper bound of 4 times the radius. However, we also see that the mean of the distribution is much closer to 0 than the maximum. Factors that play a role in this are the directedness and density of the network.

² The data was anonymised, so no individual or pair of people could be identified.

³ We used networkx for processing, geopandas for geocoding, matplotlib for plotting, osmnx for import, and scipy to compute the distance matrix.

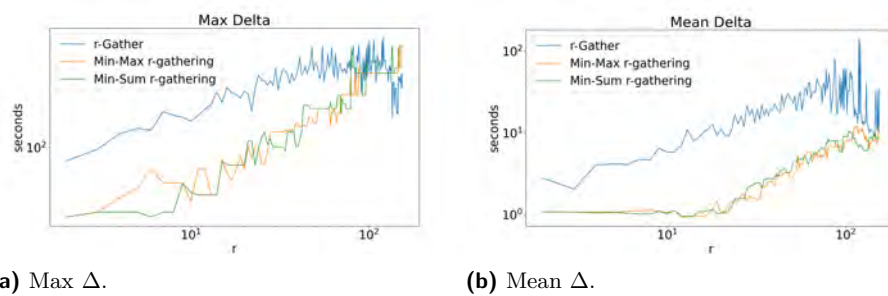


■ **Figure 3** Shown are different clusterings with the minimum capacity of $r=100$ of Bonn Ückesdorf in Germany. In general, the mean Δ is far from the maximum and clumped around 0 (blue histogram on the right side of each subfigure with a log scale). 3a shows the min-max r -gathering with the best maximum cluster size and mean time in seconds. 3b shows the min sum r -gathering. 3c shows the r -gather, which has a higher meantime because it only finds large clusters. 3d shows the clustering from Haunert et al. [5]. Here, the cluster of locations are retrieved from the buildings in OpenStreetMaps. As they grow trees until they are big enough, these clusters are connected.

We show the example of Bonn Ückesdorf, a small suburb. This allows us to compare our clusterings with the clustering of Haunert et al. [5]. Figure 3 shows the four clustering strategies. The clusters are randomly coloured, and we depict streets without inhabitants as thin grey lines. On the right of each subfigure are the histograms of the Δ of each route between different clusters. The r -gathering approaches lead to a significantly smaller mean Δ . The approximation algorithm for r -gather produces equal-sized clusters. The clusters in r -gather have the problem that they are not connected. The disconnection comes from the flow problem satisfying the minimum capacity. Here the edges can be arbitrarily distributed between the clusters when they have enough edges and their influence radius overlap.

In Figure 4, we compare Δ for different r and strategies. In this setting, r -gather gives bad results for small r but catches up for larger r . It also seems to be less stable as the other. Surprisingly, the r -gatherings stay close to each other in max and mean Δ .

All clustering strategies had runtimes from a few seconds to minutes on a city scale. The computations were done with a regular desktop pc and programmed in Python. Anonymising a route does not require much more time than a normal routing query. We need to look up the centres of the cluster of our endpoints and query the route between the centres. Finding the boundary point of the cluster on the route is straightforward, and routing to these



■ **Figure 4** The graph on the left depicts the max Δ and on the right mean Δ for different r for Bonn Ückesdorf. Both axes use a log scale.

checkpoints only needs a short-distance route query. Furthermore, routing between different cities can be done by clustering every city individually. Therefore, it is only necessary that the different clustering do not overlap, as that would break the k -anonymity.


4 Conclusion

We have developed a framework for anonymous routing that has minimal impact on travel time. We explored four minimum capacity clustering strategies and their effects on travel time. Our analysis revealed that the r -gather and min-max r -gathering strategies provided upper bounds for the maximum extra travel time. We also examined the min-sum r -gathering strategy and found that both r -gathering cluster strategies resulted in shorter mean extra travel times. In the future, we plan to use a weighted version of the clustering strategies to aggregate edges in the graph, which might lead to faster computations. Additionally, we believe that a finer subdivision of the streets could reduce extra travel time even further.

References

- 1 Gagan Aggarwal, Rina Panigrahy, Tomás Feder, Dilys Thomas, Krishnam Kenthapadi, Samir Khuller, and An Zhu. Achieving anonymity via clustering. *ACM Trans. Algorithms*, 6(3), 2010. doi:10.1145/1798596.1798602.
- 2 Amitai Armon. On min-max r -gatherings. *Theoretical Computer Science*, 412(7):573–582, 2011. Selected papers from WAOA 2007: Fifth Workshop on Approximation and Online Algorithms. doi:10.1016/j.tcs.2010.04.040.
- 3 Holger Bast, Stefan Funke, Peter Sanders, and Dominik Schultes. Fast routing in road networks with transit nodes. *Science*, 316(5824):566–566, 2007. doi:10.1126/science.1137521.
- 4 Anna Brauer, Ville Mäkinen, Axel Forsch, Juha Oksanen, and Jan-Henrik Haunert. My home is my secret: concealing sensitive locations by context-aware trajectory truncation. *International Journal of Geographical Information Science*, 36(12):2496–2524, 2022. doi:10.1080/13658816.2022.2081694.
- 5 Jan-Henrik Haunert, Daniel Schmidt, and Melanie Schmidt. Anonymization via clustering of locations in road networks. In *GIScience 2021 Short Paper Proceedings*. UC Santa Barbara: Center for Spatial Studies., 2021. doi:10.25436/E2CC7P.
- 6 Latanya Sweeney. k -anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570, 2002. doi:10.1142/S0218488502001648.
- 7 Statistische Ämter des Bundes und der Länder. Bevölkerung im 100 meter-gitter, 2018. URL: <https://www.zensus2011.de/DE/Home/Aktuelles/DemografischeGrunddaten.html?nn=559100>.

Achieving Least Relocation of Existing Facilities in Spatial Optimisation: A Bi-Objective Model

Huanfa Chen¹ ✉ 🏠 

Centre for Advanced Spatial Analysis, University College London, UK

Rongbo Xu ✉

Centre for Advanced Spatial Analysis, University College London, UK

Abstract

Spatial optimisation models have been widely used to support locational decision making of public service systems (e.g. hospitals, fire stations), such as selecting the optimal locations to maximise the coverage. These service systems are generally the product of long-term evolution, and there usually are existing facilities in the system. These existing facilities should not be neglected or relocated without careful consideration as they have financial or management implications. However, spatial optimisation models that account for the relocation or maintenance of existing facilities are understudied. In this study, we revisit a planning scenario where two objectives are adopted, including the minimum number of sites selected and the least relocation of existing facilities. We propose and discuss three different approaches that can achieve these two objectives. This model and the three approaches are applied to two case studies of optimising the retail stores in San Francisco and the large-scale COVID-19 vaccination network in England. The implications of this model and the efficiency of these approaches are discussed.

2012 ACM Subject Classification Information systems → Geographic information systems

Keywords and phrases spatial optimisation, location set cover problem, multiple objective

Digital Object Identifier 10.4230/LIPIcs.GIScience.2023.19

Category Short Paper

1 Introduction

Spatial optimisation or facility location models are aimed at siting facilities so as to provide service to demands efficiently. A range of location models have been proposed to support varying management, planning, and decision-making contexts. In particular, the location set cover problem (LSCP) [9] has been proposed for planning applications in which the fewest facilities are to be sited so as to serve all demand within the designated service response standard. The LSCP can be written as [2]:

$$\text{Minimize } \sum_{j=1}^n x_j \quad (1)$$

Subject to:

$$\sum_{j \in N_i} x_j \geq 1 \quad \forall i \quad (2)$$

$$x_j \in \{0, 1\} \quad \forall j \quad (3)$$

¹ Corresponding author



19:2 Least Relocation LSCP

Where:

$i =$	index referencing nodes of the network as demand
$j =$	index referencing nodes of the network as potential facility sites
$n =$	total number of potential sites
$S =$	maximal acceptable service distance or time standard
$d_{ij} =$	shortest distance or travel time between nodes i and j
$N_i =$	$\{j \mid d_{ij} < S\}$
$x_j =$	$\begin{cases} 1, & \text{if a facility is located at node } j \\ 0, & \text{otherwise} \end{cases}$

In applications where there are one or more existing facilities, the LSCP in its basic form as above is faced with a major problem, as it does not differentiate between sites with and without existing facilities. The scenarios where the relocation of existing facilities is concerned are common in applications. In this paper, we focus on a planning scenario where two objectives are adopted: the first one is overall efficiency, which is exactly the objective of LSCP (see Formula 1). This objective requires the least number of sites to be selected, regardless of sites with and without existing facilities. The second objective, called the least relocation of existing facilities, dictates the maximum maintenance of existing facilities (or the least relocation of existing facilities), meaning that as many existing facilities as possible should be utilised. These two objectives are not equally important and the first criterion has a higher priority than the second one.

This bi-objective problem was introduced by [8] and illustrated by a case study of optimising the locations of fire companies for the Denver fire department. This bi-objective LSCP is as follows:

$$\text{Minimize } \sum_{j=1}^n x_j \quad (4)$$

$$\text{Maximize } \sum_{j=1}^p x_j \quad (5)$$

Subject to:

$$\sum_{j \in N_i} x_j \geq 1 \quad \forall i \quad (6)$$

$$x_j \in 0, 1 \quad \forall j \quad (7)$$

In addition to the notations above, the following notations are used:

$j =$	index referencing nodes of the network as potential sites. Sites with existing facilities are indexed from 1 to p . Sites without are indexed from $p+1$ to n .
$p =$	total number of sites with existing facilities, $p \leq n$.

In the following, we present three approaches that are applicable to solve this problem.

The first approach is proposed in the paper as mentioned above [8], which combines the two objectives into a single one and transforms the bi-objective problem into a single-objective programming problem. More details of this approach can be found in [8].

The second approach is inspired by [7]. Originally, the author proposed a method to deal with sites with and without current facilities in spatial optimisation by keeping a specified number of current facilities in LSCP. Here, we extend this approach to solve the bi-objective LSCP. Specifically, this approach adds an additional constraint to LSCP that keeps a specified number (r) of current facilities and relocating others. By iterating all possible r values and solving a list of LSCP problems with different r , a pool of LSCP solutions with different r would be obtained, and then the LSCP solution with the maximum r would be the final solution to the bi-objective problem.

Third, this problem can be directly solved using a hierarchical or lexicographic method [6]. Specifically, Objective (4) is assigned with a higher priority than Objective (5), and these two objectives are optimised in priority order. This approach is incorporated in general-purpose mixed programming solvers like Gurobi [5]; however, it is not supported in others such as GLPK [4].

While these three approaches would derive optimal solutions to the bi-objective LSCP, the computing efficiency of these approaches are understudied. In the following section, we will compare these approaches using two case studies with different problem sizes.

2 Case studies

We present two case studies to compare the function and performance of the three approaches to the bi-objective LSCP. All processing and computation are conducted on a desktop MacOS 10.15.5, 2.7 GHz with 8 GBytes memory.

2.1 Case study of siting stores in San Francisco

In this case, a retail chain would like to site a number of stores in San Francisco. The primary objective is to locate stores close to population centres, which are represented by 205 census tracts in this city. In this problem, we consider a set of 16 potential store sites and set the maximum service distance to access a store on the road network as 5 kilometres. The facility-demand distance matrix was derived from ArcGIS Network Analyst extension [1]. To simulate the scenario with a set of existing facilities, we randomly chose eight sites and assumed that there were existing facilities at these sites. This bi-objective problem is formulated as below: given the existing eight stores and a set of eight potential sites, at least how many sites should be selected to site the stores to cover all populations?

2.2 Case study of COVID-19 vaccination network in England

This case study aims to optimise the COVID-19 vaccination network in England. England contains 56.6 million people in 2020, which accounts for 84.3% of the UK's population. During the COVID-19 pandemic, a COVID-19 vaccination network was built and maintained to provide vaccination to residents, and this network consisted of 1,600 vaccination centres by November 2021. The locations of these vaccination centres are likely not optimised and some centres are redundant. Therefore, we formulate the location optimisation problem of the COVID-19 vaccination network as follows: given the existing 1,600 vaccination centres and a set of 21,127 potential sites (based on locations of the Point Of Interest), at least how many sites should be selected to locate the vaccination centres to cover all populations?

The demands in this problem are the populations of each Middle Layer Super Output Area (MSOA), with population-weighted centroids of MSOAs as demand points and the population of the 2011 census as weights.

2.3 Results and discussion

The results of the two case studies are presented in Table 1. Both cases verify that these three approaches derived optimal solutions with the same number of selected sites with existing facilities and sites without. In terms of computing efficiency, in the small-size San Francisco case, the three approaches solved the location problem using less than one second, demonstrating high computing efficiency. In contrast, when solving the large-size COVID-19 vaccination case, Approach 3 significantly outperformed the other two approaches regarding the computing time.

■ **Table 1** Example of test session results.

Case study (n, p) ¹⁾	San Francisco (8, 8)	England (21127, 1600)
Approach 1 Weighted	(4, 4, 0.1s) ²⁾	(313, 107, 512m 53.9s)
Approach 2 Iterative	(4, 4, 0.2s)	(313, 107, 294m 6.5s)
Approach 3 Lexicographic	(4, 4, 0.1s)	(313,107, 54m 53.5s)

- 1) n and m represent the number of sites without and with existing facilities, respectively
- 2) the three numbers represent the number of selected sites without existing facilities, number of selected sites with existing facilities, and computing time

3 Conclusions

In this paper, we revisited a bi-objective extension of LSCP that aims to achieve two objectives simultaneously, including the minimal number of selected sites (with higher priority) and the maximal number of selected sites with existing facilities. We show that this problem can be tackled by three different approaches, using two planning cases. The results verify that these three approaches are capable of tackling this bi-objective LSCP. In terms of computing efficiency, while these approaches exhibit similar computing time in the small-size case of San Francisco, the third approach (lexicographic) shows significantly higher efficiency than the other two approaches.

This research opens up avenues for future research. First, we will attempt to analyse and understand the computational complexity of the three approaches. Second, we plan to incorporate this bi-objective LSCP into the *spopt* Python library [3], an emerging open-source project for spatial optimisation.

References

- 1 Huanfa Chen, Alan T. Murray, and Rui Jiang. Open-source approaches for location cover models: capabilities and efficiency. *Journal of Geographical Systems*, 23(3):361–380, April 2021. Publisher: Springer. doi:10.1007/s10109-021-00350-w.
- 2 Richard L. Church and Alan Murray. *Location covering models: History, applications and advancements*. Springer, New York, 2018.
- 3 Xin Feng, James D Gaboardi, Elijah Knaap, Sergio J Rey, and Ran Wei. Pysal/spopt. doi:10.5281/zenodo.4444156.

- 4 GNU Project. GLPK (GNU Linear Programming Kit), version 4.65, 2017. URL: <http://www.gnu.org/software/glpk>.
- 5 Inc. Gurobi Optimization. Gurobi Optimizer Reference Manual, 2016. URL: <http://www.gurobi.com>.
- 6 Miettinen Kaisa. *Nonlinear Multiobjective Optimization*, volume 12 of *International Series in Operations Research & Management Science*. Kluwer Academic Publishers, Boston, USA, 1999.
- 7 Alan T. Murray. Optimising the spatial location of urban fire stations. *Fire Safety Journal*, 62(PART A):64–71, November 2013. Publisher: Elsevier. doi:10.1016/j.firesaf.2013.03.002.
- 8 Donald R. Plane and Thomas E. Hendrick. Mathematical programming and the location of fire companies for the denver fire department. *Operations Research*, 25(4):563–578, August 1977. Publisher: INFORMS. doi:10.1287/opre.25.4.563.
- 9 Constantine Toregas, Ralph Swain, Charles ReVelle, and Lawrence Bergman. The location of emergency service facilities. *Operations Research*, 19(6):1363–1373, October 1971. Publisher: INFORMS. doi:10.1287/opre.19.6.1363.

Exploring Energy Deprivation Across Small Areas in England and Wales

Meixu Chen¹  

Department of Geography and Planning, University of Liverpool, UK

Alex Singleton  

Department of Geography and Planning, University of Liverpool, UK

Caitlin Robinson  

School of Geographical Sciences, University of Bristol, UK

Abstract

Building on a growing field of research on vulnerability to energy poverty, this study focused on addressing the rising energy crisis by examining the issue of energy deprivation in local areas of England and Wales. We developed a classification for energy deprivation using a clustering method to group multiple indicators across various domains. By doing this, we identify spatial disparities of energy deprivation for people living in different neighbourhoods, aiming to provide valuable insights for governments, charities and stakeholders and inform policy making and intervention.

2012 ACM Subject Classification General and reference → Measurement

Keywords and phrases energy deprivation, spatial inequality, vulnerability, geodemographics

Digital Object Identifier 10.4230/LIPICs.GIScience.2023.20

Category Short Paper

Funding This work is funded by the UK ESRC Consumer Data Research Centre (CDRC) grant reference ES/L011840/1.

Acknowledgements I want to thank colleagues from Geographic Data Science Lab for their advice on the classification labelling.

1 Introduction

The energy market experienced strain in 2021 driven in part by the rapid economic recovery following the COVID-19 pandemic. However, the situation escalated into a global energy crisis, particularly affecting Europe, when the Russian Federation militarily intervened in Ukraine in February 2022 [6]. The crisis has led to a significant increase in living costs, in particular energy costs, resulting in an estimated 6.7 million UK households experiencing energy poverty by November 2022 [8].

Energy poverty or deprivation, as defined by Bouzarovski [1], refers to the lack of access to affordable, reliable, and environmentally friendly energy services, such as heating and lighting, that are adequate in quality and safety [5, 11]. Energy deprivation negatively impacts on health, well-being, social relationships, education, and economic development [1, 7] and poses challenges to the UK government's goal of achieving net-zero greenhouse gas emissions by 2050 [3].

In response to the energy crisis and energy poverty, it is crucial to understand the spatial distribution and characteristics of energy deprivation to inform policy and practice, as well as gaining insights into broader socio-economic and political factors contributing to the

¹ mark corresponding author



20:2 Exploring Energy Deprivation Across Small Areas in England and Wales

problem. This study aims to develop a nationwide classification of energy deprivation at small area scale to improve understanding and support evidence-based decision-making by policymakers.

2 Data and Methods

■ **Table 1** Selected variables and their descriptions.

Code	Variable name	Description
V01	Efficient energy	Energy efficiency bands are rated from A (most efficient) to G (least efficient)
V02	Inefficient energy	Efficient energy refers to properties rated as band A and B
V03	Fossil fuels dependency	Inefficient energy includes properties rated as band E, F and G
V04	High CO2 emissions	Fuel type of the property belongs to one of the fossil fuels
V05	Old property	Carbon dioxide emission per square meter of the property is higher than the average
V06	New property	Properties built before 1930
V07	No central heating	Properties built after 2012
V08	Not connected to gas grid	Households with no access to central heating
V09	Prepayment electricity meters	Domestic properties not connected to the mains gas grid
V10	Renewable energy	Households with prepayment electricity meters
V11	Electricity energy	Households with renewable energy access only
V12	Age 0 to 4	Households with electricity access only
V13	Age 75 years and over	Households with young children aged four and below
V14	Lone parent with dependent children	Households with older adults aged 75 years and over
V15	Large household size	A dependent child is any person aged 0 to 15 in a household or aged 16 to 18 in full-time education and living a family with their parent or grandparent
V16	Under occupancy	More than five people living in a household
V17	Retired	Households with at least one bedroom more than required
V18	Long-term sick and disabled	Economically inactive population that aged 16 years and over who did not have a job between 15 March to 21 March 2021 and had not looked for work between 22 February to 21 March 2021 or could not start work within two weeks.
V19	Looking after home or family	Household owns all of the accommodation
V20	Detached house or bungalow	Property is not attached to another property but can be attached to a garage
V21	Semi-detached house or bungalow	Property is joined to another property by a common wall that they share
V22	Terraced	Property located between two other properties and shares two common walls or is part of a terraced development but only shares one common wall.
V23	Flat	Property in a purpose-built block of flats or tenement
V24	Shared houses	Property part of a converted or shared house, including bedsits
V25	Owens outright	Household owns all of the accommodation
V26	Owens with mortgage or shared ownership	Household owns with a mortgage or loan, or part-owned on a shared ownership
V27	Socially rented	Property rented through a local council or housing association
V28	Privately rented	Property rented through a private landlord or letting agent
V29	More income on energy cost	Percentage of household net income spent on the electricity and gas bills
V30	Elementary occupation	Persons aged 16 years and over who do elementary job as their main occupation
V31	Unpaid care with more than 20 hours	Persons that look after, give help or support to anyone who has long-term physical or mental ill-health conditions, illness or problems related to old age
V32	Unemployment	Persons that have not worked in the last 12 months and never worked
V33	Part-time employment	Persons who worked 30 hours or less (including paid and unpaid overtime) a week before the Census
V34	Full-time students	Economically inactive full-time students
V35	Ethnic minority	Persons who are not English, Welsh, Scottish, Northern Irish, or British
V36	Universal credit	Persons who are not English, Welsh, Scottish, Northern Irish, or British
		A single payment for each household to help with living costs for those on a low income or out of work

Data are collected from multiple data sources in England and Wales, including Department for Levelling Up, Housing and Communities (DLUHC), Department for Business, Energy and Industrial Strategy (BEIS), Department for Work and Pensions (DWP), and the 2021 UK Census that are available at the property, postcode, and the Lower Layer Super Output Areas (LSOAs) levels, respectively. LSOAs were created as a geographical structure to enhance the collection and presentation of detailed statistical data for small areas in England and Wales. To ensure effective and timely representation, all data are accessed from the most recent years since 2018. 2021 Census LSOA geography in England and Wales are used as the unified spatial granularity to link with data at diverse geographical scales.

We follow a typically geodemographic classification method framework to build an energy deprivation classification. First, a list of variables is selected based on the large amounts of review of energy vulnerability and poverty [1, 7, 9, 10, 12, 13]. 36 variables are selected to reflect the energy deprivation and can be summarised into five domains: energy efficiency, energy access, energy demand and service, housing and financial vulnerability. Figure 1 depicts the chosen variables and their descriptions. All variables are measured using percentages to reduce the potential data bias of various estimation size available at individuals, households, or properties.

Prior to clustering, transformation and standardisation are conducted to enable equal variable contribution and more interpretable results. Additionally, correlation analysis is implemented to avoid certain types of variables with a high degree of association skewing

the cluster result. We exclude variables that exhibit correlation coefficient values larger than 0.8 (either positively or negatively) with more than one other variables. Five variables, specifically high carbon dioxide emissions, prepayment electricity meter, under occupancy, retired, and universal credit, are identified and excluded. Lastly, a widely used k-means clustering method [2, 4, 14] is conducted to group all LSOAs in England and Wales. To determine the optimal number of clusters (k), a Clustergram is utilised to help identify the point of diminishing returns, where increasing the number of clusters does not significantly improve the clustering quality. K=6 finally generates robust results after multiple iterations.

3 Results and Discussions

Figure 1 displays the spatial disparities of six groups of energy deprivation at LSOAs in England and Wales, representing Energy Efficient Suburbs, Energy Periphery, Energy Density, Energy Inefficiency, Energy Constraints, and Energy Precarity (Group A to F). For better interpretation, we calculate index scores of each variable and create Figure 2 to help us explain the energy deprivation characteristics for each group.

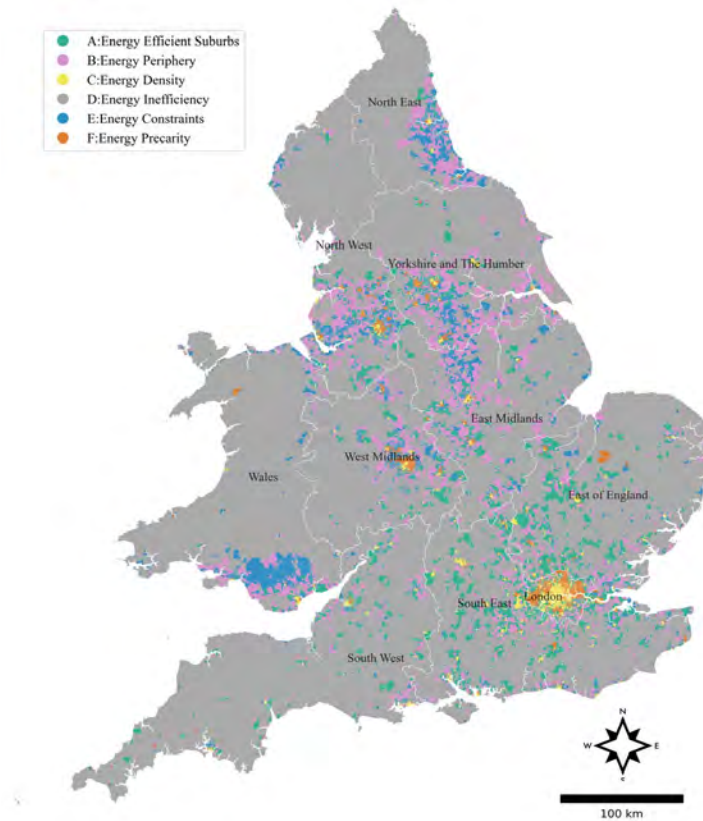
Residents of Group A, Energy Efficient Suburbs, typically live in relatively new houses with the highest energy efficiency and lowest carbon footprints compared to other groups. They tend to own these houses financed using a mortgage, loan or shared ownership scheme. Properties in the group are typically well-connected to the gas grid. There is a higher proportion of families have very young children below four years old. The group is found throughout suburban areas in England and Wales, especially in the southeast and southwest regions of England.

Group B, Energy Periphery, is characterised by residents of retirement age and who are mostly white British, own their detached or semi-detached property either outright, or with a mortgage or via shared ownership. Properties are typically well serviced by energy, including central heating and are well-connected to mains gas. However, properties tend to be under-occupied and hence their occupants consume more energy than they might in smaller homes. This group is pervasive in urban outskirts, and towns close to cities.

For Group C, Energy Density, many individuals are economically inactive full-time students and ethnic minority, concentrating in high-density neighbourhoods of privately rented flats or shared houses. They rely heavily on electricity as gas grid access is often limited. Additionally, residents may reside in either older properties without central heating or properties with high energy efficiency ratings A or B. The group is concentrated in the city centres of England and Wales.

Neighbourhoods classified in group D, Energy Inefficiency, are predominately located across rural parts of England and Wales. Residents are typically older, retired and tend to live in detached houses that they own outright. Properties are typically built before 1930 and some lack a gas grid connection due to their rurality. Most properties have low energy efficiency, leading to higher carbon dioxide emissions per square meter. Some properties only use renewable energy resulting in lower carbon footprints.

Group E, Energy Constraints, is typified by residents who are white British and have constrained access to energy services, predominantly concentrated at urban edges and suburbia of the north and midlands of England and southern Wales. Residents typically reside in rented semi-detached or terraced social housing, and are employed in elementary occupations. They often receive welfare payments to cover essential living costs. Many energy precarious households are lone parents with dependent children below four years old, or have



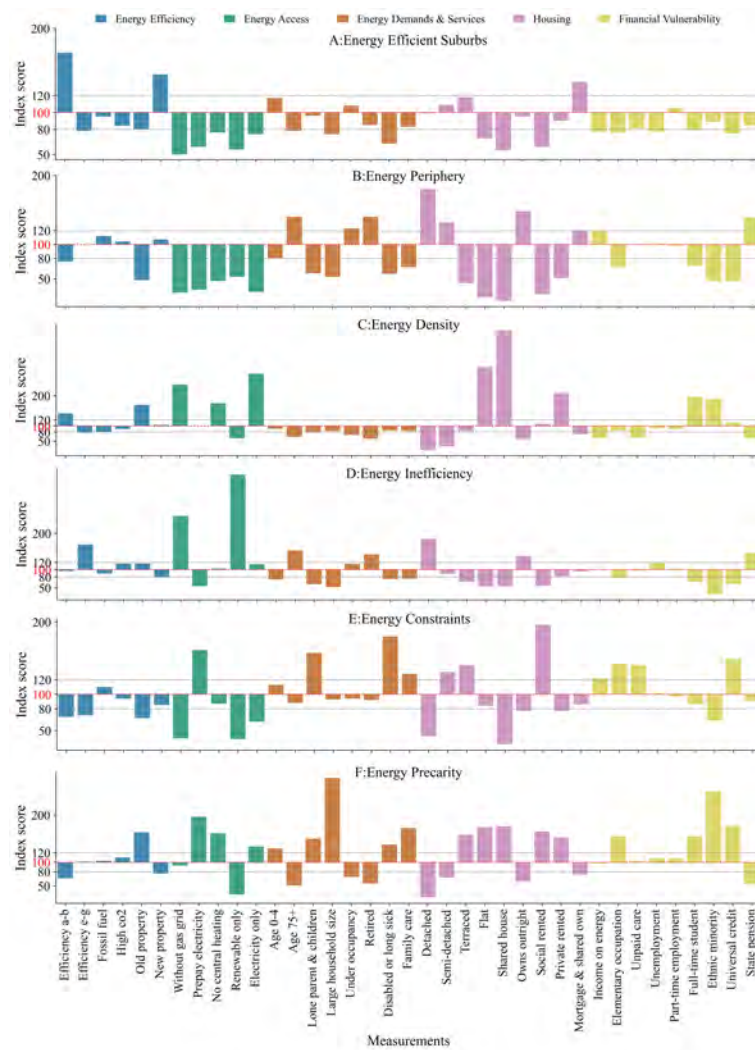
■ **Figure 1** Spatial patterns of energy deprivation in England and Wales at LSOA scale.

residents who provide unpaid care or have long-term sick or disability. Residents often use prepayment electricity meters to manage their energy bills, which are more expensive than other payment ways, and thus, higher proportions of their income are spent on energy.

Residents of Group F, Energy Precarity, are the most energy-deprived compared to other groups. Neighbourhoods offer a mix of rented terraced, flats and shared older properties, that often have constrained energy access, including no central heating, dependent on electricity only and prepayment electricity meters. These low-income areas have a high proportion of ethnic minorities, lone parent households and dependent children. They are more likely to live in overcrowded properties. The group is prevalent in outer parts and less desirable neighbourhoods of cities and towns.

4 Conclusion

This research collected and measured multiple indicators related to energy deprivation. By examining the spatial distribution and contextual characteristics of cluster results, we identify the most and least energy deprived areas in England and Wales and the characteristics of individuals living in those areas. Some future works are required to mitigate the limitations. First, there is no best method to determine the number of optimal k and k -means clustering has some embedded limitations such as sensitive to outliers, exploration with other methods may assist in more accuracy and reliability in methodology. Furthermore, the selection of variables may not cover all factors that influence the energy poverty of households due to



■ **Figure 2** Index score for each group of energy deprivation.

the data limitation of small area statistics. Further survey data can be used via small area micro-simulation to supplement more variables for the classification. Lastly, this study mainly focus on the description of energy deprivation classification, further policy implications should be provided for future work.

References

- 1 Stefan Bouzarovski and Saska Petrova. A global perspective on domestic energy deprivation: Overcoming the energy poverty-fuel poverty binary. *Energy Research and Social Science*, 2015. doi:10.1016/j.erss.2015.06.007.
- 2 Meixu Chen, Dominik Fahrner, Daniel Arribas-Bel, and Francisco Rowe. A reproducible notebook to acquire, process and analyse satellite imagery. *REGION*, 2020. doi:10.18335/region.v7i2.295.
- 3 Department for Business Energy and Industrial Strategy. Net Zero Strategy: Build Back Greener, 2021. URL: <https://www.gov.uk/government/publications/net-zero-strategy>.

- 4 Christopher G. Gale, Alexander D. Singleton, Andrew G. Bates, and Paul A. Longley. Creating the 2011 area classification for output areas (2011 OAC). *Journal of Spatial Information Science*, 12(2016):1–27, 2016. doi:10.5311/JOSIS.2016.12.232.
- 5 Mikel González-Eguino. Energy poverty: An overview, 2015. doi:10.1016/j.rser.2015.03.013.
- 6 IEA. Global Energy Crisis, 2022. URL: <https://www.iea.org/topics/global-energy-crisis>.
- 7 Christine Liddell and Chris Morris. Fuel poverty and human health: A review of recent evidence. *Energy Policy*, 2010. doi:10.1016/j.enpol.2010.01.037.
- 8 National Energy Action. Energy Crisis, 2022. URL: <https://www.nea.org.uk/energy-crisis/>.
- 9 Office for National Statistics. Age of the property is the biggest single factor in energy efficiency of homes, 2022. URL: <https://www.ons.gov.uk/peoplepopulationandcommunity/housing/articles/ageofthepropertyisthebiggestsinglefactorinenergyefficiencyofhomes/2021-11-01>.
- 10 Saska Petrova. Encountering energy precarity: Geographies of fuel poverty among young adults in the UK. *Transactions of the Institute of British Geographers*, 2018. doi:10.1111/tran.12196.
- 11 K. N. Reddy, David Bloom, Anita Kaniz, and Mehdi Zaidi. Energy and Social Issues. *Energy*, 2000.
- 12 Caitlin Robinson, Stefan Bouzarovski, and Sarah Lindley. “Getting the measure of fuel poverty”: The geography of fuel poverty indicators in England. *Energy Research and Social Science*, 2018. doi:10.1016/j.erss.2017.09.035.
- 13 Caitlin Robinson, Sarah Lindley, and Stefan Bouzarovski. The Spatially Varying Components of Vulnerability to Energy Poverty. *Annals of the American Association of Geographers*, 2019. doi:10.1080/24694452.2018.1562872.
- 14 Alex Singleton, Alexandros Alexiou, and Rahul Savani. Mapping the geodemographics of digital inequality in Great Britain: An integration of machine learning into small area estimation. *Computers, Environment and Urban Systems*, 2020. doi:10.1016/j.compenvurbsys.2020.101486.

Using the Dynamic Microsimulation MINOS to Evidence the Effect of Energy Crisis Income Support Policy

Robert Clay ¹   

University of Leeds, UK

Luke Archer   

University of Leeds, UK

Alison Heppenstall   

School of Political and Social Sciences, MRC/CSO Social and Public Health Sciences Unit, University of Glasgow, UK

Nik Lomax   

School of Geography, University of Leeds, UK

Abstract

Rates of anxiety and depression are increasing due to financial stress caused by energy pricing with over half of UK homes unable to afford comfortable heating. UK Government policies to address this energy crisis have been implemented with limited evidence and substantial criticism. This paper applies the dynamic microsimulation MINOS, which utilises longitudinal Understanding Society data, to evidence change in mental well-being under the Energy Price Cap Guarantee and Energy Bill Support Scheme Policies. Results demonstrate an overall improvement in Short Form 12 Mental Component Score (SF12-MCS) both on aggregate and over data zone spatial areas for the Glasgow City region compared with a baseline of no policy intervention. This is work in progress and discussion highlights potential future work in other energy policy areas, such as Net Zero.

2012 ACM Subject Classification Applied computing → Sociology

Keywords and phrases Dynamic Microsimulation, Mental Health, Energy Poverty

Digital Object Identifier 10.4230/LIPIcs.GIScience.2023.21

Category Short Paper

Supplementary Material *Software (Source Code)*: https://github.com/Leeds-MRG/Minos/tree/244_gis, archived at `swh:1:dir:be994021b118b5533e0d37815ce7cf198ed048c2`

Funding This work was supported by the UK Prevention Research Partnership (MR/S037578/1, Meier), which is funded by the British Heart Foundation, Cancer Research UK, Chief Scientist Office of the Scottish Government Health and Social Care Directorates, Engineering and Physical Sciences Research Council, Economic and Social Research Council, Health and Social Care Research and Development Division (Welsh Government), Medical Research Council, National Institute for Health Research, Natural Environment Research Council, Public Health Agency (Northern Ireland), The Health Foundation and Wellcome.

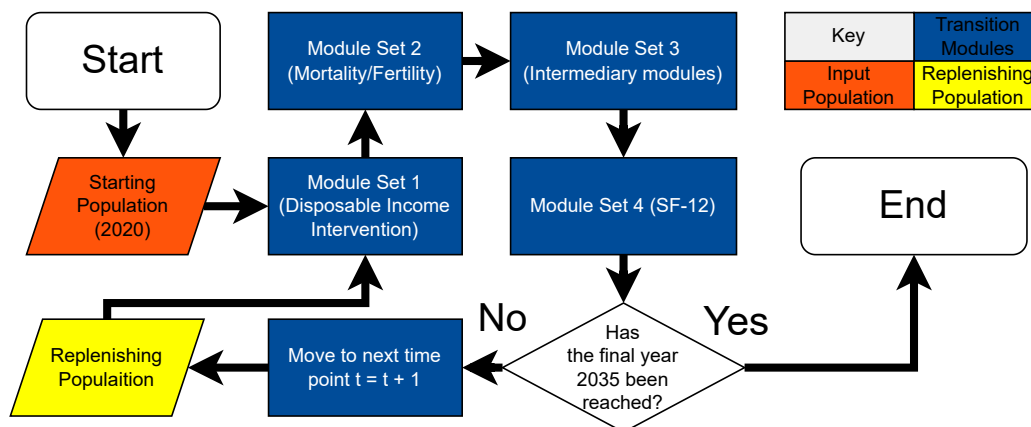
1 Introduction

United Kingdom energy prices have tripled in just five years, impacting on the cost of living via fuel, food, and other expenses [3]. Household income is being squeezed to the point that more than half of UK Homes [11] are now in energy poverty and can no longer afford

¹ Corresponding Author



21:2 Applying MINOS to Energy Crisis Policy



■ **Figure 1** MINOS model flow chart. Outlines three main model stages for initialisation of 2020 population, propagation through a series of transition models and population replenishment for 15 years until 2035.

comfortable heating. Literature [2] suggests this is having a drastic effect on mental health as financial stressors increase incidence of anxiety and depression. The UK Government is implementing a number of policies to try and protect household disposable income and public health. Two well reported examples [5] are the Energy Price Cap Guarantee (EPCG) which places a temporary cap on energy prices and the Energy Bill Support Scheme (EBSS) where a flat £400 rebate is provided to all households, with more going to vulnerable subgroups such as pensioners. These policies have been widely criticised as insufficient and based on limited evidence [3, 9, 7]. Work is needed to quantify the effect of the energy crisis on mental health and any alleviating effects from both real EBSS policy and any hypothetical alternatives [9].

One methodology to synthesise policy evidence is dynamic microsimulation [10]. A population of individual households is generated, either synthetically or from survey data, and propagated forwards in time under some transition probability mechanics and policy interventions. Dynamic microsimulation can be used to quickly generate long term evidence over a suite of multiple policies repeatedly to quantify uncertainty and provide rich individual level data. This approach can be readily used by policymakers for evidence-based decision making and has seen broad application particularly in economics and health [10, 6]. This paper applies the dynamic microsimulation MINOS to estimate how the EPCG and EBSS policies have affected UK mental health using the Short Form 12 Mental Component Summary (SF12-MCS) [12] as the outcome of interest. Section 2 outlines the overall MINOS model structure, data sources, transition probability models and intervention scenarios. Section 3 presents results demonstrating change in mental well-being due to both policies at aggregate level and spatially over Data Zones (DZs) for the Glasgow City region. Finally section 4 summarises findings, limitations, and potential future work.

2 Data and Methods

MINOS is built using standard dynamic microsimulation design [10] as a first order discrete time Markov model. UK population households transition forwards one year at a time using a series of transition models that estimate new state at time $t + 1$ only using current time t information. MINOS is built using a flexible modular design allowing for associative representation of causal systems pathways between income and mental well-being [8] given

available data. MINOS is completely open source and written in the R. and python languages using the vivarium framework [1]. An overview of the MINOS life cycle is given in Figure 1. The first stage initialises a population of UK households by importing either real or synthetic panel data. This population serves as initial conditions and is the same for every model run. The second stage of MINOS is a series of transition probability modules. Each module evolves some subset of individual attributes forward one year in time. The order in which these modules are run is important and done in four sets. The first set only updates household disposable income. All policy interventions are parameterised as change in household disposable income, and placing this module first allows change to propagate through the rest of the system immediately. The second set contains birth and death modules to ensure only those who are alive are intervened upon. The third set contains a number of intermediary modules that are influenced by change in disposable income and subsequently influence mental well-being. The final fourth set contains only a mental well-being module that estimates the desired health outcome given change in all other areas. After running each module set, the third stage of MINOS is population replenishment where a new batch of households is added to the model to maintain population size. Stages two and three are run indefinitely until the desired simulation horizon year 2035. At this point any post-hoc analysis can be performed to estimate life trajectories of households. A summary of the transition probability models used is given in table 1. Each model used depends primarily on outcome data type. Full module data tables, predictors, and model coefficients are available on GitHub². MINOS uses the Understanding Society (US) dataset [4] as its primary data source. US provides 11 annual cohorts from 2009-2020 containing hundreds of individual and household attributes pertaining to health, employment, and demographics. US data are used both directly as input population data for MINOS and to calibrate statistical models. Preprocessing is applied to US data to correct missing data and improve readability resulting in $n = 15192$ individual observations for the 2020 starting year dataset. Additional ONS data are used to calculate mortality and fertility rates using the NEWETHPOP project [13]. Key variables in the US dataset are household disposable income and Short Form 12 Mental Component Score (SF12-MCS). Household disposable income is defined as money after taxes and bills a household can spend on other needs. It is a continuous variable that is heavily right skewed with a median income of £1300 per month. SF12-MCS is a widely used metric of mental well-being that ranges from 0 – 100 with higher values indicating better well-being. It is approximately Gaussian distributed with mean 50 and variance 10^2 . In order to produce spatially disaggregated results, we draw on a synthetic population which combines US data with Census data to derive Data Zone level estimates for Scotland [14]. Results use a 10% sample scale for the Glasgow City region with $n = 227589$ individuals over 747 data zones each containing 500 – 1000 people and mapped in Figure 2b. Glasgow was chosen for its mixed socioeconomic demographics and an existing Scottish evidence base for comparison [3].

Three scenarios are applied to the Glasgow City region. The first baseline scenario changes nothing about the UK population. Energy prices start low at 2018 levels and never increase. This is a useful benchmark with which to compare other policies. The second policy is the Energy Price Cap Guarantee (EPCG) scenario, in which energy prices do increase and are capped by the UK government. Data suggests [11] energy prices increased by 240% from 2020 to 2023. High energy prices are sustained beyond 2023 to assess the impact to mental well-being. The final scenario implements both the EPCG and EBSS policies together to assess any further change in well-being. The EBSS provides a £400 base lump sum to all

² https://github.com/Leeds-MRG/Minos/tree/244_gis

21:4 Applying MINOS to Energy Crisis Policy

■ **Table 1** Table of transition probability models used in MINOS.

Module Name (Module Set)	Outcome Variable	Variable Type	Model Used
Household disposable income (1)	Monthly household disposable income (£s)	Continuous	Ordinary Least Squares
Mortality (2)	Is the subject alive? (yes/no)	Binary	Rate Tables
Fertility (2)	Has subject given birth in last year? (yes/no)	Binary	Rate Tables
Housing quality (3)	Household quality composite (1-3 Likert)	Ordinal	Cumulative Link Model
Neighbourhood safety (3)	Neighbourhood composite (1-3 Likert)	Ordinal	Cumulative Link Model
Loneliness (3)	Is subject lonely? (1-3 Likert)	Ordinal	Cumulative Link Model
Nutrition (3)	How many fruit and vegetables per week? 0+	Continuous	Ordinary Least Squares
Tobacco (3)	Cigarettes smoked per week 0+	Continuous	Zero Inflated Poisson
Mental Well-Being (SF12) (4)	SF12 Mental Component Summary score (0-100)	Continuous	Ordinary Least Squares

households and further income to households that include pensioners, universal credit, long term sick/disability, and council tax bands A-D. Each of these groups receives £650, £150, £300, and £150 additional monthly income respectively [5, 3]. These changes are applied directly to household disposable income and are capped at an upper limit of £0 matching real policy that only returns money if energy is used. Each of these three policies is run 100 times through MINOS.

3 Results

Results estimate changes in well-being for the Glasgow City population at an aggregate level. For each of the MINOS runs the mean SF12-MCS value is recorded giving a sample of 100 means. These are then used to estimate confidence intervals for the overall SF12-MCS mean. Values are then scaled to estimate the percentage change in well-being versus baseline for EPCG and EBSS policies in Figure 2a. The EPCG policy sees an approximate 0.25% decrease in SF12-MCS score per year for the population by 2026. EPSS policy does improve mental well-being vs EPCG alone but is not enough to bring SF12-MCS score back to pre-2018 levels. This finding matches other literature suggesting the energy crisis has been detrimental to mental well-being [3] but underestimates the effect size for reasons mentioned in final discussion. SF12-MCS values were also aggregated spatially over Scottish Data Zones. Starting with the mean by data zone for each of the 100 model/simulation runs, and taking the grand mean again results in a scalar change in SF12-MCS for each data zone that can be mapped. Figure 2b shows these values for the year 2025 comparing difference in SF12-MCS for the EPCG vs EBSS scenarios. The EBSS scenario shows an improvement in mental well-being across all areas. The areas that see the least improvement appear to be the most economically affluent. Unsurprisingly these households are largely insensitive to energy price

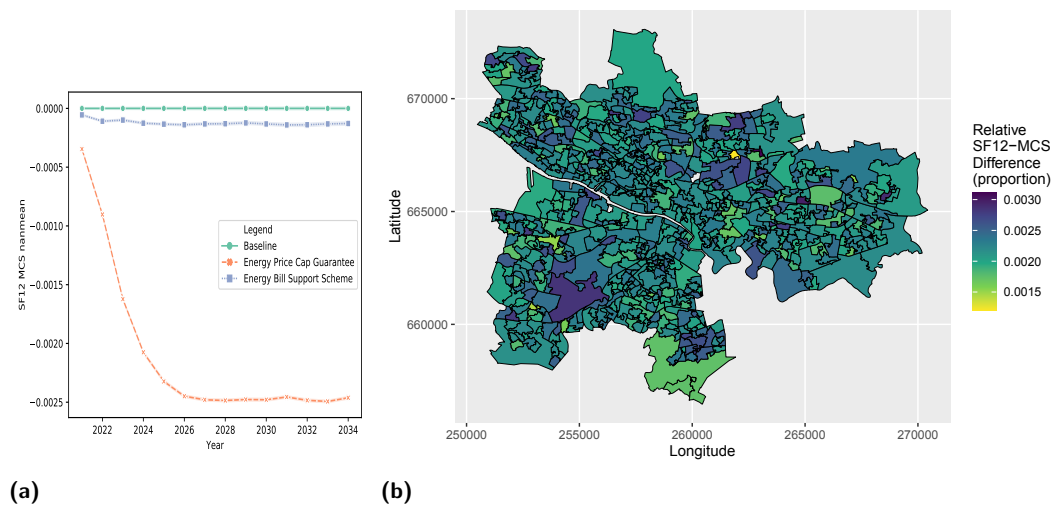


Figure 2 Energy interventions for Glasgow: (a) Lineplot showing relative change in SF12-MCS score for EPSS and EPCG vs baseline. (b) Map of the Glasgow City region comparing EPCG and EBSS policies for the year 2025. Comparison with the Scottish Index of Multiple Deprivation (<https://simd.scot>) shows a strong correlation (0.61) between deprived LSOAs and greater benefit from the EBSS scheme. The Springburn (262000,668000), Castlemilk (260000,660000), and Pollok (254000,662000) regions see the most benefit. Note two LSOAs are missing with white colouration as map geometry is from 2011 and they no longer exist in 2020.

increases. Lower-middle class areas of Glasgow (north east, south east) appear to benefit the most from the EBSS policy. Poorer areas see most of their disposable income lost to the energy crisis and preventing this improves mental well-being [9]. Further investigation is needed to disaggregate this map and identify vulnerable households that are seeing less improvement under the EBSS policy [9] and personalise policy strategy.

4 Discussion

This paper has applied the dynamic microsimulation framework MINOS to the Energy Bill Support Scheme policy in the UK to estimate its impact on household disposable income and mental well-being. Results show significant overall improvement in well-being when implementing EBSS policy, but not enough to return mental well-being to 2018 baseline levels. Spatial analysis suggests lower income households benefit the most from EBSS policy; however more analysis is required to determine which vulnerable households need more assistance.

This work is limited primarily due to the Understanding Society data source. Yearly interval data are not granular enough to capture mental health change that can occur at a much finer timescale (e.g. days). There is also missing data for food and vehicle fuel expenditure, this makes it difficult to accurately estimate changes in household disposable income. Comparison with other research suggests the effect sizes for the EBSS are underestimated [3].

Application of the Energy Bill Support Scheme policy demonstrates the utility of the MINOS framework. Initial future work will elaborate on the methods used in MINOS particularly for transition probabilities and validation techniques using literature and online documentation. MINOS can readily be further developed to simulate spatially distributed policies that target inequities such as prepayment meters, the influence of place (e.g. rural

locations) and additional cost of living impacts such as increase in rent [3]. In combination with other Government targets [9, 5] such as social housing, energy efficiency and Net Zero, MINOS has the potential to be developed into a pragmatic tool for future crisis policy that is able to balance the preservation of economy and health as well as protecting vulnerable households.

References

- 1 Institute for Health Metrics and Evaluation, University of Washington, Seattle, USA . Vivarium dynamic microsimulation framework. <https://github.com/ihmeuw/vivarium>. Accessed: 2023-05-22.
- 2 Virginia Ballesteros-Arjona et al. What are the effects of energy poverty and interventions to ameliorate it on people’s health and well-being?: A scoping review with an equity lens. *Energy Research & Social Science*, 87:102456, 2022.
- 3 Philip Broadbent, Rachel Thomson, Daniel Kopasker, Gerry McCartney, Petra Meier, Matteo Richiardi, Martin McKee, and Srinivasa Vittal Katikireddi. The public health implications of the cost-of-living crisis: outlining mechanisms and modelling consequences. *The Lancet Regional Health–Europe*, 27, 2023.
- 4 Nick Buck and Stephanie McFall. Understanding society: design overview. *Longitudinal and Life Course Studies*, 3(1):5–17, 2011.
- 5 Daniel Harari, Brigid Francis-Devine, Paul Bolton, and Matthew Keep. Rising cost of living in the uk. *London: House of Commons Library* <https://commonslibrary.parliament.uk/research-briefings/cbp-9428>, 2022.
- 6 Andreas Höhn, Jonathan Stokes, Roxana Pollack, Jennifer Boyd, Cristina Chueca Del Cerro, Corinna Elsenbroich, Alison Heppenstall, Annika Hjelmkog, Elizabeth Inyang, Daniel Kopasker, et al. Systems science methods in public health: what can they contribute to our understanding of and response to the cost-of-living crisis? *J Epidemiol Community Health*, 2023.
- 7 Nada Khan. The cost of living crisis: how can we tackle fuel poverty and food insecurity in practice? *British Journal of General Practice*, 72(720):330–331, 2022.
- 8 Petra Meier, Robin Purshouse, Marion Bain, Clare Bamba, Richard Bentall, Mark Birkin, John Brazier, Alan Brennan, Mark Bryan, Julian Cox, et al. The sipher consortium: introducing the new uk hub for systems science in public health and health economic research. *Wellcome open research*, 4(174):174, 2019.
- 9 Amy Norman and Scott Corfe. Energy bill support—designing policies to support british households in an age of high prices. *Social Market Foundation*, 2022.
- 10 Martin Spielauer, Gerard Thomas Horvath, and Marian Fink. microwelt: A dynamic microsimulation model for the study of welfare transfer flows in ageing societies from a comparative welfare state perspective. Technical report, WIFO Working Papers, 2020.
- 11 UK Government Department for Business, Energy, and Industrial Strategy (BEIS). Quarterly energy prices. <https://www.gov.uk/government/collections/quarterly-energy-prices>. Accessed: 2023-05-21.
- 12 John E Ware Jr, Mark Kosinski, and Susan D Keller. A 12-item short-form health survey: construction of scales and preliminary tests of reliability and validity. *Medical care*, pages 220–233, 1996.
- 13 Pia Wohland, Phil Rees, Paul Norman, Nikolas Lomax, and Stephen Clark. Newethpop-ethnic population projections for uk local areas 2011-2061. *UK Data Service*, 2022.
- 14 Guoqiang Wu, Alison Heppenstall, Petra Meier, Robin Purshouse, and Nik Lomax. A synthetic population dataset for estimating small area health and socio-economic outcomes in great britain. *Scientific Data*, 9(1):19, 2022.

Multiscale Spatially and Temporally Varying Coefficient Modelling Using a Geographic and Temporal Gaussian Process GAM (GTGP-GAM)

Alexis Comber¹  

School of Geography, University of Leeds, UK

Paul Harris  

Sustainable Agriculture Sciences, Rothamsted Research, Harpenden, UK

Chris Brunsdon  

National Centre for Geocomputation, National University of Ireland, Maynooth, Ireland

Abstract

The paper develops a novel approach to spatially and temporally varying coefficient (STVC) modelling, using Generalised Additive Models (GAMs) with Gaussian Process (GP) splines parameterised with location and time variables - a Geographic and Temporal Gaussian Process GAM (GTGP-GAM). This was applied to a Mongolian livestock case study and different forms of GTGP splines were evaluated in which space and time were combined or treated separately. A single 3-D spline with rescaled temporal and spatial attributes resulted in the best model under an assumption that for spatial and temporal processes interact a case studies with a sufficiently large spatial extent is needed. A fully tuned model was then created and the spline smoothing parameters were shown to indicate the degree of variation in covariate spatio-temporal interactions with the target variable.

2012 ACM Subject Classification Information systems → Spatial-temporal systems

Keywords and phrases Spatial Analysis, Spatiotemporal Analysis

Digital Object Identifier 10.4230/LIPIcs.GIScience.2023.22

Category Short Paper

Funding This work was supported by the JSPS BRIDGE fellowship No. 220305.

1 Introduction

This paper describes a novel approach for spatially and temporally varying coefficient (STVC) modelling. It extends Geographical Gaussian Process GAMs (GGP-GAM) to include GP splines parameterised space and time. GGP-GAMs have been shown to be more accurate than Multiscale Geographically Weighted Regression (MGWR)[1] the effective SVC brand leader. GGP-GAMs explicitly accommodate process spatial heterogeneity and provide an alternative to assumptions of stationarity[4]. STVCs extend this to the temporal dimension.

Generalized Additive Models (GAMs) are general in that they can handle outputs with many types of distributions and not just linear relationships, polynomial or not. They are additive and because they generate multiple model terms which are added together to generate predictions. The advantages of a GAM-based approach to SVC and STVC modelling are because GAMs are flexible and able to handle different types of response[6, 2]. This is due to their additive nature which combines multiple sub-models, and the modelling of non-linear relationships using splines, the building blocks of GAMs. Splines are combination of functions (*basis functions*) which may be single or multi-dimensional, each of which is

¹ Corresponding author



assigned a coefficient, which are combined to generate \hat{y} and in this way complex relationships are modelled in GAMs. Splines parameterised with location form the basis of SVC modelling with GGP-GAMs and here this is extended to include time within the splines: the geographic and temporal Gaussian process GAM (GTGP-GAM).

The inclusion of temporal data can provide insight on the spatial process dynamics and a number of approaches that include time in spatial regressions exist. However, with the exception of GTWR, most of these are concerned with capturing autocorrelation effects, rather than relationship heterogeneity. This paper extends GGP-GAMs to the temporal dimension. Temporal process are well described by GPs. However a key methodological consideration is *how* space and time should be analysed together. This paper uses a national case study to investigate the relative benefits of combining space and time into a single 3D GP spline against treating them separately in a 2D + 1D approach.

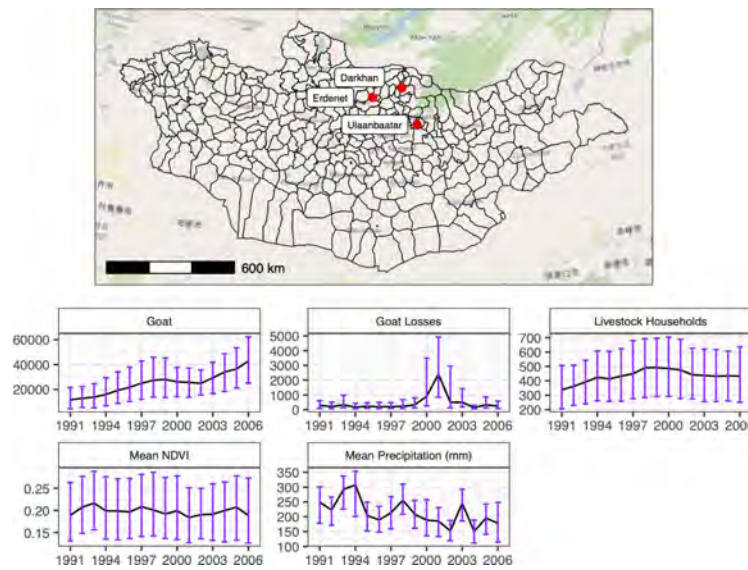
2 Case study

A case study of national data of livestock in Mongolia as reported in [5] was analysed. This reports livestock totals, here focussing on goats, over 341 soums (second-level administrative units) from 1991-2006 and these were considered to be a function of annual mean normalised difference vegetation index (NDVI), annual mean rainfall, the number of households working with livestock and the number of reported animal losses in the previous year. The choice of historical loss data and household working with livestock, as explanatory variables was because livestock losses have been found to play a critical role in livestock decisions and viability [3]. The environmental variables reflect the changes in biomass and their drivers over time. Figure 1 shows the spatial context of the soums and the trends over time of the variables. These indicate a steady increase in cattle numbers, which is associated with increased meat consumption (anecdotally increasingly concentrated around Ulaanbaatar). The goat losses indicate the dzud period 2001–2002 which are extreme weather events associated with deep snow, severe cold and conditions that make foraging difficult and results in livestock deaths. The number of households associated with livestock production increases and levels off as livestock management becomes more concentrated. The the median of mean monthly NDVI is relatively stable, and mean monthly precipitation shows some fluctuation.

3 Results

A key issue in space time analysis is to determine whether observation spatial and temporal variables should be handled separately or together, ie whether their covariances are separable or non-separable. One way to approach this is to construct models and compare their performance through some measure of model fit such as AIC, and prediction accuracy such as mean absolute error (MAE). Here the aim was to construct models of goat numbers. These have a classic Poisson distribution. One option is to construct Poisson regression models and another is to transform the response variable and fit Gaussian regression models. A square root transform of the goat counts was undertaken here. Four GAM models were constructed with GP splines parameterised with location and temporal data separately and together, using normalised (z-scores) spatial and temporal data and the original spatial and temporal data. For each of these the model MAE and AIC are summarised in Table 1, with the “best” model determined from the AIC measure.

The best performing model was one which combined space and time, with normalised space and time variables. This indicates the interaction between the space and time effects in this case and their lack of independence. This is perhaps not surprising due to the large



■ **Figure 1** The Soums in Momgolia (n = 341) with the 3 largest cities (top), and the trends in the median values of the variables, with upper and lower quartiles indicated (bottom).

■ **Table 1** Summaries of the model predictive performance and fits, with separate and combined GP splines for locational and temporal variables, and with normalised and un-normalized data.

Splines	Normalized.Data	MAE	AIC
Separate	No	9593.17	-1800.22
Separate	Yes	9608.42	-1941.88
Combined	No	9405.98	-1812.12
Combined	Yes	9776.30	-1972.61

spatial extent of the case study and thus the spatial variation of the drivers of local goat numbers are likely to vary over space and over time. That is, different effects are more likely to be experienced in different places at any given time and the pattern of these is more likely to change over time. Also it is important to note that although the results indicate that a better performing model is obtained by combining space and time, this is not to suggest that no distinct effects occur. Rather that the uncertainty in calibrating a more complex model with 3D splines leads to more reliable prediction.

The GTGP-GAM models were created using default parameters for the `gam` function in the `mgcv` package. The convergence of the spline smoothness optimisation of the best performing model was examined in detail, and specifically the effect for the number of knots (k) used to construct the spline basis dimensions. Investigation indicated that k was potentially too low, with the effective degrees of freedom (EDF) for some splines close to k . The models was tuned by increasing k to 400 resulting in improved model fits (AIC) and convergence of the GP splines as the high k values ensured sufficient degrees of freedom in the splines.

The fixed parametric coefficient estimates are shown in Table 2. These show significant intercepts and generally insignificant covariates (except for mean NDVI in the Mongolian case study). The smooth terms for the combined spatial and temporal GP splines (ie the STVCs) are summarised in Table 3. The full set of coefficients are not printed because there many coefficients for each spline, one for each basis function. The `edf` (effective degrees of freedom) summarises the complexity of the spline smooths, with an `edf` value of 1 indicating

■ **Table 2** The GTGP-GAM fixed parametric coefficients and their global significance.

Covariate	Estimate	Std. Error	t-value	p-value
Intercept	113.795	2.597	43.813	0.000
Goat Losses	0.209	0.413	0.507	0.612
Livestock Households	1.209	1.244	0.972	0.331
Mean Precipitation (mm)	-0.738	2.101	-0.351	0.725
Mean NDVI	-83.217	103.519	-0.804	0.422

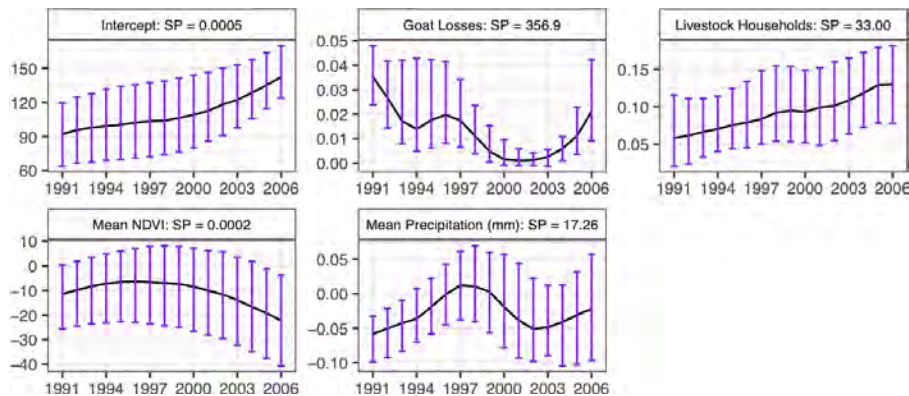
■ **Table 3** The GTGP-GAM smooth terms of the tuned model.

GTGP Spline	edf	Ref.df	F	p-value
s(X,Y,Ti):Intercept	10.688	11.515	22.090	0.000
s(X,Y,Ti):Goat Losses	206.717	248.831	2.704	0.000
s(X,Y,Ti):Livestock Households	220.063	256.766	7.805	0.000
s(X,Y,Ti):Mean Precipitation (mm)	143.082	168.939	4.311	0.000
s(X,Y,Ti):Mean NDVI	5.272	6.302	0.566	0.761

a straight line, 2 a quadratic curve etc. Higher **edf** values indicate increasing non-linearity in the relationship between the covariate and the response. The p-values relate to splines / smooths defined over these, and their significance can be interpreted as indicating whether they vary locally over space and time combined (i.e. spatio-temporally). That is, covariates with insignificant p-values (i.e. Mean NDVI) still have an effect, but these effects do not vary locally. In contrast to the fixed parametric coefficients, all are significant. That is, their relationship with the target variable y varies locally over space and serially over time with different temporal effects in different places.

It is possible to extract the spatially and temporally varying coefficient estimates. These describe how the relationship between y and the covariates varies over space and time. It is instructive to examine these alongside the GTGP-GAM smooth terms or smoothing parameters (SPs). The SPs indicate the scale of the relative spatial-temporal variation of the interaction between each covariate and the response. These and summaries of the STVCs over time are shown in Figure 2. This plots the median coefficient estimates for each year, describing how the coefficient estimates vary over time, with their variation over space summarised in their inter-quartile range (IQR). It shows that:

- The Intercept steadily increases over time and the the IQR gradually narrows towards the end of the sequence. It has a low SP indicating stable spatial relationships over time.
- The association of Goat Losses with goat numbers decreases to 2002 and then increases to 2006. However, the IQR shows high variation over time indicating, narrowing to 2002 and the increasing in later years. This has a high spline SP value, indicating a strong spatially and temporally varying relationship with the target variable.
- The relationship of Livestock Households with goat numbers steadily increases over time. This may indicate the impact of an increasing concentration of livestock within fewer households. The variation (IQR) over space also remains relative stable, with a small degree of variation, reflected in the moderate spline SP value.
- Mean NDVI and Mean Precipitation are both mostly negative in the association with goat numbers, with Mean NDVI decreasing in later years and Mean Precipitation increasing to zero and the decreasing before increasing in later years. There is more spatial variation over time in Mean Precipitation than Mean NDVI, as reflected in their SP values.



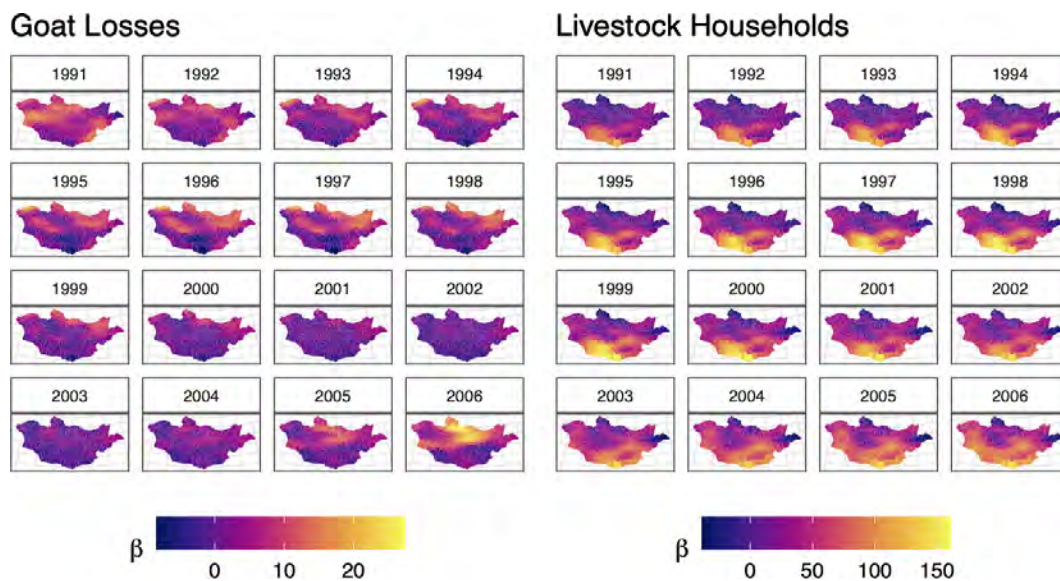
■ **Figure 2** The temporal trends in the median values of the coefficient estimates over time (with upper and lower quartiles) indicated, with spline smoothing parameters (SP).

In summary, Figure 2 shows the temporal trend of the relationships between the covariates and the target variable. The changes in the IQRs over time indicate whether the relationship and thus the process is changing spatially as well as temporally. This interaction between space and time in the STVCs is also reflected in the spline SP values.

It is also possible to confirm this interpretation of the spline SPs and the nature of the STVCs they indicate: broadly, larger SPs indicate greater spatio-temporal interaction of the covariate with the target variable. Figure 3 compares the coefficient estimates for the Goat Losses (high SP) and Mean NDVI (low SP). For the Goat Losses, this shows how the spatial relationship between the covariate and the target variable (Goat population) changes over time, with stronger relationship in and around the major population centres noticeable in 2006, for example. By contrast Mean NDVI has a spatially varying relationship with the target variable but this does not change over time.

4 Brief Discussion and Conclusions

The paper explores STVC modelling through the application of GAMs with GP spline parameterised with location and time variables. Here these were combined into a single 3-dimensional spline for each predictor variable, under the assumption that the case study extent was sufficiently large for an assumption of the geographic and temporal process interacting over space and time to hold. In this model the temporal trends in the relationship between the predictor variable and the target variable were allowed to vary with location. The paper demonstrates STVC modelling using GAMs with GP spline parameterised with location and time variables. Different GP spline compositions were explored to determine whether space and time should be treated separately or assumed to interact. In this case, exploring a national case study, the best fitting model was found to be one in which space and time measurements were re-scaled to z-scores and combined in 3-dimensional GP spline. This reflected *a priori* assumption that spatial and temporal processes would interact for case studies with a sufficiently large spatial extent, an assumption that proved to be true in this case. In other situations this might not be the case. Case studies with smaller spatial extents, or indeed with shorter runs over time, may require location and time to be treated separately, with separate GP splines for each predictor variable. In this situation, the assumption would be that the spatial and temporal trends in the data and their relationship with the outcome do not interact, and that any changes in the relationship with target variable over time would





■ **Figure 3** Examples of STVC estimates for Goat Losses and Livestock Households over a 16 year time period.

be independent of changes in location. These models exclude the possibility of different temporal trends in different locations. It also reflected an assumption that AIC as a measure of model fit and parsimony identified the “best” model. Other investigations (not reported here) has suggested that BICs (Bayesian Information Criteria) may be more appropriate for investigating and comparing space-time interacted models with independent space and time effects. Future work will explore these issues. The model was then tuned with large number of knots, allowing sufficient degrees of freedom for the model parameters. The relative values of the tuned model smoothing parameters provided an indication of the variation in the spatial and temporal interactions of the covariates with the target variable. Summaries over time of the median values of the coefficient estimates demonstrated the temporal trends, and the spatially varying nature of these was suggested by the interquartile ranges of these. These were confirmed by the smoothing parameters and through visual exploration.



References

- 1 Alexis Comber, Paul Harris, and Chris Brunsdon. Multiscale spatially varying coefficient modelling using a geographical gaussian process gam. *International Journal of Geographical Information Science*, submitted.
- 2 Ludwig Fahrmeir, Thomas Kneib, Stefan Lang, and Brian D Marx. Regression models. In *Regression*, pages 23–84. Springer, 2021.
- 3 John McPeak. Confronting the risk of asset loss: What role do livestock transfers in northern kenya play? *Journal of Development Economics*, 81(2):415–437, 2006.
- 4 Stan Openshaw. Developing gis-relevant zone-based spatial analysis methods. *Spatial analysis: modelling in a GIS environment*, pages 55–73, 1996.
- 5 Narumasa Tsutsumida, Paul Harris, and Alexis Comber. The application of a geographically weighted principal component analysis for exploring twenty-three years of goat population change across mongolia. *Annals of the American Association of Geographers*, 107(5):1060–1074, 2017.
- 6 Simon N Wood. *Generalized additive models: an introduction with R*. Chapman Hall/CRC, 2006.

Does Generalisation Matters in Pan-Scalar Maps?

Azelle Courtial¹  

LASTIG, Univ Gustave Eiffel, IGN-ENSG, F-77420 Champs-sur-Marne, France

Guillaume Touya  

LASTIG, Univ Gustave Eiffel, IGN-ENSG, F-77420 Champs-sur-Marne, France

Abstract

Maps and their usage have widely evolved recently, to become more and more interactive, multi-scale and accessible. However, the design of maps did not change so much, leading to the following two problems: (1) in theory, it is not formalised how to create a good map in this context, (2) in practice, the most used maps are not good considering the quality criteria defined for the classical (static) maps. Therefore, it is necessary to question the usefulness of these principles in this new context. In this article, we focus on the role of cartographic generalisation in maps where one can easily zoom in and out to make information accessible. We draw up a list of hypotheses on the role of generalisation for pan-scalar maps, based on both a deductive approach (the role of map generalisation is deduced from a review of human-maps interactions), and an inductive approach (observation of maps with diverse qualities). Then, we discuss how these hypotheses might be experimentally verified.

2012 ACM Subject Classification Applied computing → Cartography

Keywords and phrases map generalisation, cartography, pan-scalar map, multi-scale map, spatial cognition

Digital Object Identifier 10.4230/LIPICs.GIScience.2023.23

Category Short Paper

Funding This project has received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (grant agreement No. 101003012 - LostInZoom).

1 Introduction

Map generalisation (MG) is the process of deriving a map at a certain scale from detailed geographic information. This process is an important step in designing a static map at a specific scale. However, most maps used today are not static paper maps at a specific scale or even a stack of independent static maps but pan-scalar maps [7]. “Pan-scalar map” refers to interactive zoomable applications comprised of numerous maps of a particular area at different zoom levels (i.e. scales), and we assume that the generalisation of such maps involves new challenges compared to traditional map generalisation. Today, a broad audience can easily access many maps from a computer or mobile device. In parallel with the multiplication of cartographic media, the time spent on each has decreased and the user expects ever faster and more easily accessible geo-information. We could think that a simple response to this need would be to use more and more generalised maps, as MG reduces the level of detail, hierarchises and simplifies the information in each view. In practice, we rather observe the contrary: Google Maps, for instance, is globally chosen by users although it includes many views with a lousy generalisation, an ill-adapted level of detail, and remaining conflicts. This observation raises the question if MG still matters when you can just zoom in to see more details or zoom out to see less.

¹ Corresponding author



The goal of this article is to refine the usefulness of MG in the current context, by examining how the quality of MG may affect users during the exploration of pan-scalar maps. We first review how pan-scalar maps are used, then we review the graphical consequence of a bad generalisation and their impact on map usage and finally, we draw up a list of hypotheses on map generalisation usefulness and discuss how they can be verified.

2 How do people use maps?

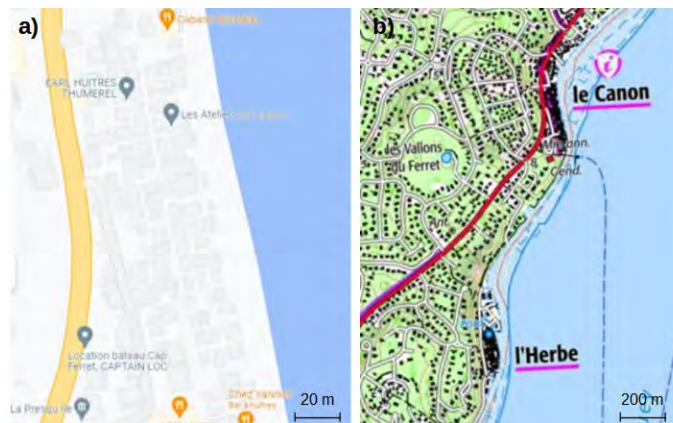
Maps are often defined as communication models to convey geographic information, and the most general usage of maps is the acquisition of spatial knowledge [11]. Understanding map reading strategies for spatial knowledge acquisition in pan-scalar maps involves both the understanding of the reading strategy for each view [1] and the understanding of the navigation strategy. In this context, an exploration is a set of successive views and transitions triggered by interaction (e.g. zoom-in, zoom-out, pan) [14]. For each new view of the exploration, a reconciliation is performed to associate the information of this view with the one from the previous views (present in short-term memory), and the user's previous knowledge (present in long-term memory). This step is both achieved from information observed pre-attentively and attentively and the feeling of disorientation in a pan-scalar map is provoked by contradictory or missing information during reconciliation. Thus, MG can affect disorientation by its impact on the representation of entities used for reconciliation.

Moreover, the typification of cartographic usage is an important subject in spatial cognition and the usage of the map may widely affect the exploration strategy and thus the role of generalisation. For instance, thematic maps represent and locate phenomena in relation to the environment; and their generalisation should highlight spatial relations between the phenomena and their context and their different magnitudes across scale [5]. Consequently, the exploration should contain a few interactions and longer views. On the contrary, route planning is often made from topographic maps and involves many interactions: zoom out on the complete path; zoom in on arrival, departure place and complex intersection; pan along the route etc. In this case, the role of generalisation is to preserve road hierarchy and connections necessary for route planning and avoid disorientation at each interaction.

3 Map generalisation and pan-scalar maps

Map generalisation is the adaptation of the level of detail at a certain scale. This reduction of the level of detail should enhance legibility while preserving the main structures and patterns at the target scale. The role of MG in the usability of static maps has been formalised years ago, and to our knowledge, it was never updated for pan-scalar maps. Some studies already demonstrate the impacts of some map design choices (related to MG) on the efficiency of pan-scalar map exploration (e.g. landmark visualisation [4]), while others focus on the adaptation of legibility thresholds used in MG for screen display [12].

Currently, most pan scalar maps are created with generalisation strategies that do not differ so much from those used for papers map: (1) the range of possible scales is split into a defined number of representations (2) each is generalised independently (with a variable quality) (3) intermediate views are derived by enlarging the closest representation. This process does not take into account pan-scalar map specifics. Indeed, it aims to make each view legible not to optimise navigation between scales. Recently, there is some focus on smooth navigation [6, 15]. However, the role and usefulness of generalisation have not been investigated in a pan-scalar environment. In the next paragraphs, we review the possible effects of a bad MG in this context.



■ **Figure 1** Example of conflict that can remain after a bad generalisation. a) overlaps between buildings and roads disconnect the road network b) overlaps between buildings.

The impact of remaining graphical conflicts. First, map generalisation is often viewed as a method to resolve graphical conflicts occurring when cartographic information is rendered. A conflict is when a constraint cannot be respected (e.g. two objects overlap or an object is too small, see Figure 1). In a pan-scalar map, contrary to a static map, when a conflict occurs the user can access the hidden information by just zooming in until the conflicts disappear and thus it has less impact on the usability of the map. However, more interactions are required which is not practical, and some information theoretically visible at a certain scale can be hidden (e.g. in Figure 1.b, the two cities cannot be compared in one view). This practical question of quantity of interaction conduces to more cognitive load and more opportunity for disorientation. It may also affect the trust of the user in the map and map provider. Would you trust a map with many unsolved conflicts?

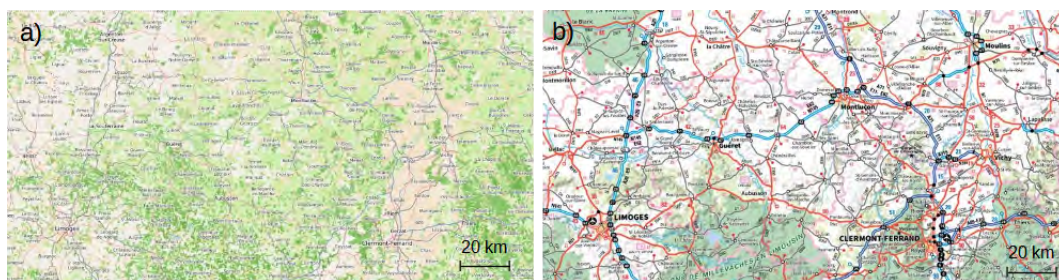
The impact of missing information. A bad generalisation may involve a loss of information: in some map views, entities are removed to avoid conflicts while it would be possible and relevant to make this information appear without conflicts with a relevant generalisation. The absence of information can be of several types: (1) the entity is not represented (2) a geometric part of the entity is missing, e.g. a notch or a meander is removed, which prevents entity recognition (3) the semantic granularity of the entity is not sufficient, e.g. in Figure 2.a. the green space is not specifically symbolized as a stadium. On one hand, such problems may hide some phenomena or information, and affect the decision-making made from the map. For instance, could you find sports facilities in Figure 2? On the other hand, it affects the exploration of the map, as some views/scales are not useful by themselves, they are just intermediate steps in the exploration, requiring more interaction overall. It may produce more disorientation and less confidence for users (1) if expected landmarks are not present, or (2) if you need to zoom too much to see the missing information.

The impact of clutter. Contrary to the previous point, a bad MG can also produce too much and too complex information. Clutter is a notion of image complexity, cluttered images tend to increase the cognitive load [8], and we observe a similar impact of clutter for map images [9]. To better understand the effect of clutter on pan-scalar map usage we can make an analogy with a messy and cluttered room, where you have to search for a particular object; even with the ability to zoom in and out, it would be easier to find the object if the room

23:4 Does Generalisation Matters in Pan-Scalar Maps?



■ **Figure 2** Illustration of missing information. a) Google map only represents the roads and green areas in this view. b) OSM depicts more information.



■ **Figure 3** An area mapped at a regional scale to illustrate the absence of hierarchy: a) OSM: very few entities are salient, and no structure stands out. b) IGN map with more hierarchy.

were organised. In a similar way memorising the position of an object is easier, and the user may experience less disorientation as the landmarks are easily identifiable and memorable between scales. Finally, we could hypothesise that cluttered maps are less accessible to the general public and require a higher level of expertise. Indeed, an expert may dedicate more time to reading the map, already knows some reading keys etc. For instance, geological maps are often cluttered, but still useful for experts.

The impact of the absence of hierarchy. Finally, map generalisation provides a hierarchy of information. Indeed the goal is not only to reduce information density and level of detail but also to highlight the main information and preserve and enhance structures and patterns. With static maps, this is mostly made by a gradual use of visual variables (e.g. size, colour) to improve salience (e.g. toponyms for cities appear with a size relative to their importance; important roads appear larger and with a brighter colour than a local road). Hierarchy is used to structure and or partition the space in a view, thus its absence makes that each entity is presented with the same level of importance. It may lead to a sense of disorientation when there is no salient landmark (see Figure 3) or when an important pattern cannot be extracted (e.g. Paris ring road does not stand out from the road network in OSM). Further than the disorientation, it is really difficult to memorise or to project yourself into a map without a hierarchy of information. Moreover, in a pan-scalar map a second factor of hierarchy exists and can be employed: the scale of appearance [7]. For now, the hierarchy in the common pan-scalar map is also unsatisfactory as structures do not appear clearly: for instance, important toponyms appear too late to avoid placement conflict, which causes confusion and cognitive load [6].

4 Measuring the effects of a good map generalisation

In the previous section, we identified several effects of map generalisation on pan-scalar map use. We review here these elements and try to identify how to measure their impact via a user test where users are faced with maps with variable quality of generalisation.

H1. Practicality. is characterised directly by the number and the range of interactions that compose an exploration. This is simple to measure and can be measured from whatever task.

H2. Disorientation. is a feeling, it can be measured by asking the participant (either with a questionnaire or think-aloud). Both strategies have drawbacks: the feeling is instantaneous so a questionnaire afterwards could be ill-adapted; a think-aloud strategy is shown to distract users from the exploration and combines very badly with eye-tracking. Then, disorientation can be indirectly measured via induced behaviours: a disoriented user may search for landmarks and thus make large saccades across the map, or zoom out strongly. Disorientation may occur in most of the tasks but it can be a tenuous phenomenon compared to all other cognitive process in play during a pan-scalar map exploration.

H3. Cognitive load. directly degrades user attention and may cause a decrease in task performance. It can also be indirectly observed by an increased number of eye blink [3]. To measure a cognitive load the task proposed to the user must be complex enough and repeated enough to cause this cognitive load.

H4. Memorisation. is the quantity and quality of remaining information after an exploration. It is commonly tested with two types of tasks: recall and recognition [2].

H5. Confidence. is the feeling of user self-confidence and confidence in the map producer during exploration and decision-making. To our knowledge, the only way to measure it is using a self-report questionnaire.

H6. Accessibility. is the level of map reading skill necessary to understand and make good decisions from the map. Measuring map reading skills is challenging. Commonly it is associated with the sense of direction, which can be estimated via self-report measures [10] or via spatial perception, mental rotation and spatial visualisation tests [13]. This skill varies in an important range in a population, thus even without a reliable estimation, it is possible to verify this hypothesis with the variability of user response for a task: the variability might be smaller for a map with good generalisation (even users with low map reading skills might succeed as well as experts). The number of participants required to observe such a variation might be important.

5 Conclusion: Does generalisation matters in pan-scalar maps?

We identified six hypotheses on the usefulness of map generalisation for pan-scalar maps: practicality, disorientation, cognitive load, memorisation, confidence and accessibility. Experimental verification of these hypotheses is crucial to design improved pan-scalar maps. However, it is a significant research challenge due to the variety of map usage, and incompatible measures. We propose the following plan: 1) Practicality and confidence can be investigated using a widely distributed study that measures user interaction and feelings for

realistic map usage. 2) Disorientation and cognitive load are studied in an in-situ survey using eye tracking. 3) Memorisation and accessibility are verified with long sessions on a specific population with various expertise levels. These three surveys are essential to discern the role played by map generalisation in the optimal use of pan-scalar maps and to guide future map generalisation research towards usability rather than tradition.

References

- 1 Marketa Beitlova, Stanislav Popelka, Martin Konopka, and Karel Macků. Verification of Cartographic Communication Models Using Detection of Map Reading Strategies Based on Eye Movement Recording. *The Cartographic Journal*, pages 1–20, 2023.
- 2 Anne-Kathrin Bestgen, Dennis Edler, Christina Müller, Patrick Schulze, Frank Dickmann, and Lars Kuchinke. Where Is It (in the Map)? Recall and Recognition of Spatial Information. *Cartographica: The International Journal for Geographic Information and Geovisualization*, 52(1):80–97, 2017. Publisher: University of Toronto Press. doi:10.3138/cart.52.1.3636.
- 3 Bingjie Cheng, Enru Lin, Anna Wunderlich, Klaus Gramann, and Sara Fabrikant. Using spontaneous eye blink-related brain activity to investigate cognitive load during mobile map-assisted navigation. *Frontiers in Neuroscience*, 17, 2023. doi:10.3389/fnins.2023.1024583.
- 4 Bingjie Cheng, Anna Wunderlich, Klaus Gramann, Enru Lin, and Sara Fabrikant. The effect of landmark visualization in mobile maps on brain activity during navigation: A virtual reality study. *Frontiers in Virtual Reality*, 3:981625, 2022. doi:10.3389/frvir.2022.981625.
- 5 C Duchêne. Making a map from “thematically multi-sourced data”: the potential of making inter-layers spatial relations explicit. In *17th ICA Workshop on Generalisation and Multiple Representation*, page 8, 2014.
- 6 Marion Dumont, Guillaume Touya, and Cécile Duchêne. Automated Generalisation of Intermediate Levels in a Multi-Scale Pyramid. In *18th ICA Workshop on Map Generalisation and Multiple Representation*, Rio de Janeiro, Brazil, 2015.
- 7 Maïeul Gruget, Guillaume Touya, and Ian Muehlenhaus. Missing the city for buildings? A critical review of pan-scalar map generalization and design in contemporary zoomable maps. *International Journal of Cartography*, pages 1–31, 2023.
- 8 Simon Harper, Eleni Michailidou, and Robert Stevens. Toward a Definition of Visual Complexity as an Implicit Measure of Cognitive Load. *TAP*, 6, 2009. doi:10.1145/1498700.1498704.
- 9 Lars Harrie, Hanna Stigmar, and Milan Djordjevic. Analytical Estimation of Map Readability. *ISPRS International Journal of Geo-Information*, 4(2):418–446, 2015.
- 10 M Hegarty. Development of a self-report measure of environmental spatial ability. *Intelligence*, 30(5):425–447, 2002. doi:10.1016/S0160-2896(02)00116-2.
- 11 Alexander Kent. Form Follows Feedback: Rethinking Cartographic Communication. *Westminster Papers in Communication and Culture*, 13:96–112, 2018. doi:10.16997/wpcc.296.
- 12 Florian Ledermann. The Effect of Display Pixel Density on the Minimum Legible Size of Fundamental Cartographic Symbols. *The Cartographic Journal*, 58(4):314–328, 2021. Publisher: Taylor & Francis. doi:10.1080/00087041.2022.2055938.
- 13 Marcia C. Linn and Anne C. Petersen. Emergence and Characterization of Sex Differences in Spatial Ability: A Meta-Analysis. *Child Development*, 56(6):1479–1498, 1985. Publisher: Wiley, Society for Research in Child Development. doi:10.2307/1130467.
- 14 Guillaume Touya, Maïeul Gruget, and Ian Muehlenhaus. Where Am I Now? Modelling Disorientation in Pan-Scalar Maps. *ISPRS International Journal of Geo-Information*, 12:62, 2023. doi:10.3390/ijgi12020062.
- 15 Peter van Oosterom, Martijn Meijers, Jantien Stoter, and Radan Šuba. Data Structures for Continuous Generalisation: tGAP and SSC. In Dirk Burghardt, Cécile Duchêne, and William Mackaness, editors, *Abstracting Geographic Information in a Data Rich World: Methodologies and Applications of Map Generalisation*, Lecture Notes in Geoinformation and Cartography, pages 83–117. Springer International Publishing, Cham, 2014.

Understanding People’s Perceptions of Their Liveable Neighbourhoods: A Case Study of East Bristol

Elisa Covato ✉

School of Computer Science and Creative Technologies,
University of the West of England, Bristol, UK

Shelan Jeawak ✉

School of Computer Science and Creative Technologies,
University of the West of England, Bristol, UK

Abstract

Liveable neighbourhoods are urban planning initiatives that aim to improve the quality of residential areas. In this paper, we focus on the East Bristol Liveable Neighbourhood (EBLN) to understand people’s perceptions of their neighbourhood’s urban reality. We analyse the opinions of citizens collected through the project, by examining their sentiments, the urban subjects they consider, and the language used to express their opinions. The findings of this study offer initial indications to inform urban planning processes and facilitate effective decision-making.

2012 ACM Subject Classification Information systems → Data analytics; Computing methodologies → Visual analytics

Keywords and phrases Urban analytics, liveable neighbourhoods, public participation geographic information system, citizen co-design, spatio-textual data, sentiment analysis, language analysis

Digital Object Identifier 10.4230/LIPICs.GIScience.2023.24

Category Short Paper

Acknowledgements We would like to thank Bristol City Council, UK, for providing the data used in this research.

1 Introduction

The term liveability has been used in various studies and at different levels of granularity ranging from individuals, neighbourhoods, and countries. It has also been used in multiple disciplines, such as geography, ecology, and urban planning [9]. Liveable Neighbourhoods (LNs) are fine-grained people-centred urban planning units with the goal of improving overall liveability. LNs aim to integrate various services and facilities in residential areas, aiming to create safe, healthy, inclusive, accessible, and attractive environments [5]. Public engagement in designing changes in their local community to meet local needs, known as co-design, plays a vital role in the development and implementation of LNs.

Understanding people’s opinions toward their neighbourhoods is crucial for informed decision-making. Researchers have employed public participation geographic information systems (PPGIS) to examine local views for urban planning and decision-making research [3, 1]. They used PPGIS to collect and analyse public perceptions across diverse landscape types and scales, with examples of application in national park planning [2] and urban planning [4]. All these studies often relied on face-to-face surveys and interviews to collect peoples’ opinions [7, 6], and comments were mostly analysed manually with respect to qualitative evaluation. However, these traditional methods are often work-intensive, and limited in sample size. To overcome these limitations, many projects are trying to use online neighbourhood reviews, that allow larger sample sizes and broader geographic coverage.



© Elisa Covato and Shelan Jeawak;
licensed under Creative Commons License CC-BY 4.0

12th International Conference on Geographic Information Science (GIScience 2023).

Editors: Roger Beecham, Jed A. Long, Dianna Smith, Qunshan Zhao, and Sarah Wise; Article No. 24; pp. 24:1–24:6

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

■ **Table 1** Example of responses in the EBLN dataset.

Sentiment	Positive	Negative
Subjects	Trees and greenery on street, Street trees and planting	Walking, Crossings
Reasons	Pleasant	Not pedestrian friendly, Difficult to cross the street
Suggestions	Slow down traffic	Add crossing, Safer junction for walking and cycling
Comments	The street planters have reduced traffic speed and made the street ‘greener’. Something similar could be done in other locations within the project area and in traffic displacement areas outside the project area	Hard to cross here - there is a traffic island slightly above this point but often want to cross lower down and it’s hard to do so as the road is busy with two lanes of fast traffic.

These typically combine numeric ratings and textual comments. However, the challenges here are in analysing such a large number of data and efficiently extracting meaningful knowledge [6].

Another group of researchers tried to use geo-tagged social media as a mirror to view public perceptions and opinions of their living environment. For example, social media data have been examined to explore people’s sentiments [10], emotions [11], satisfaction [12], and attitudes [8] toward their living area. Despite their significant findings, social media data are generally very noisy and require extensive preprocessing before use.

East Bristol Liveable Neighbourhood (EBLN) project¹ is a pilot study based on online surveys. The project aims to work with people who live, work, study, and travel through East Bristol, UK, to design people-friendly, safe, quiet, and healthy streets. It has been designed in partnership with the community as part of a co-design phase of the project which will help to shape permanent solutions.

With this work, we aim to analyse EBLN data to: (1) Understand citizen sentiment toward their living environment; (2) Investigate citizen choices of urban subjects and their mutual relations; And (3) Analyse citizen comments with respect to their sentiments.

The remainder of this paper is organised as follows. Section 2 introduces the EBLN data. Section 3 provides a detailed discussion of our analysis and results. Finally, Section 4 summarises our conclusions and plans for future work.

2 Data and study area

The survey data used in this study were collected between January and March 2022 by Bristol City Council, UK. People living, working, and travelling to or through the survey area (Figure 1b) were asked to express their views using an online interactive map². Respondents could drop a point on the map, and were then asked to:

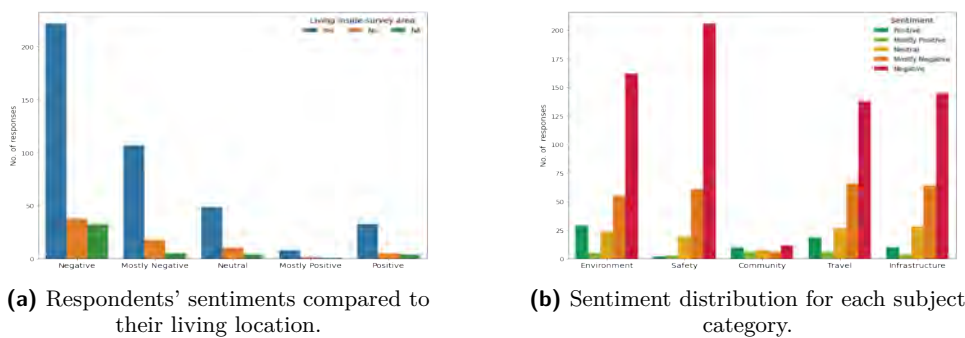
- Express their feeling by selecting one of five *sentiments*, ranging from negative to positive.
- Optionally, leave a *comment* using a free-text box.
- Optionally, select one or more *subjects* related to the comment, *reasons* for the sentiment expressed, and *suggestions* to improve the area.

¹ <https://eastbristolliveableneighbourhoods.commonplace.is>

² <https://eastbristolliveableneighbourhoods.commonplace.is/map/map>



■ **Figure 1** Geographical distribution of sentiments within and outside the East Bristol Liveable Neighbourhood survey area.



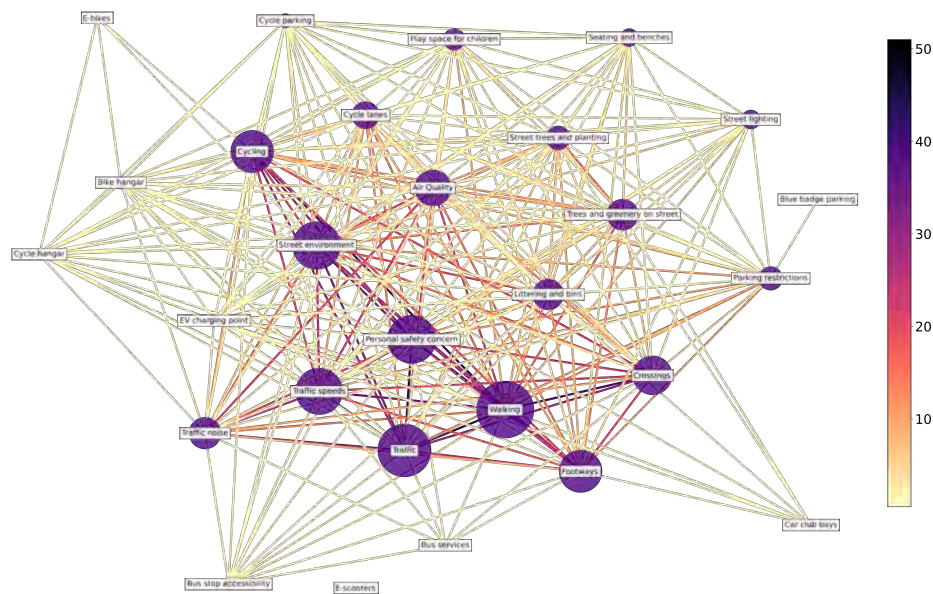
■ **Figure 2** Sentiment analysis based on respondents' living location and subject categories.

An example of responses is shown in Table 1. As Figure 1a shows, some comments refer to locations outside the survey area. Nonetheless, we have decided to include these data points in our analysis, since our final goal is to gain insights into the language citizens use to describe the urban environment around them. The dataset used comprises 540 geo-located, sentiment-based entries, of which 91% contain textual comments and subject labels. In this preliminary study of the EBLN data, we have limited our analysis to the free-text comments along with their corresponding sentiments. We have also focused on understanding the co-occurrence of urban subjects selected within the survey, as well as their relation to respondents' sentiments.

3 Analysis and results

3.1 Geographical spread and frequency of sentiments

Our initial investigation concentrated on analysing the sentiments expressed by the respondents. The primary objective was to investigate the geographical distribution of sentiments, and how they relate to whether the respondents live within the survey area or not. The findings revealed that a substantial portion of the respondents resides within the survey area (Figure 2a), with a noteworthy proportion of the sentiments expressed being characterized as negative. This highlights how PPGIS participants living in a study area are more vested in decisions regarding their community than respondents less connected to the area [1]. Due to the co-design nature of the project, it is not surprising that respondents tended to emphasize the negative aspects of their neighbourhood.

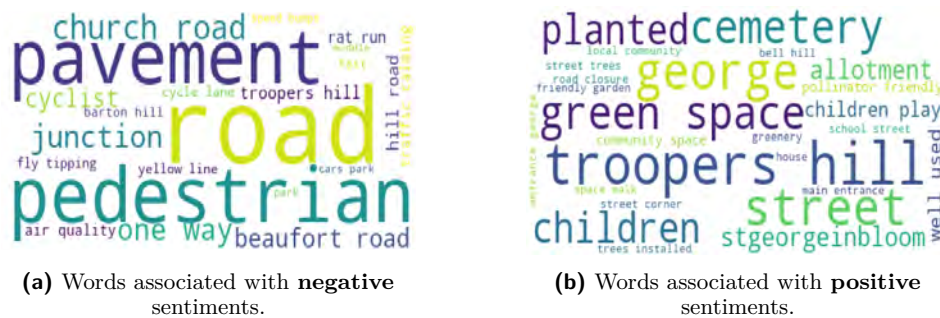


■ **Figure 3** Subjects frequency and co-occurrence. The edge colour scale represents the co-occurrence rate.

The maps in Figure 1 show that most of the negative comments are concentrated along main roads and junctions. Conversely, areas characterized by green spaces tend to display more positive-related comments. Given the aim of the EBLN project is to improve the urban environment of the neighbourhood, it is expected that the majority of comments are pinned to roads. It is worth noticing that further statistical analysis of the data revealed no significant correlation between the sentiments expressed by the respondents and the land cover and type characteristics within the study area.

3.2 Co-occurrence of subjects in the responses

In our analysis, we have identified a total of 28 distinct subjects that the respondents selected to categorize their responses. As Table 1 shows, some comments have multiple subjects associated. The top 3 most selected are: *Walking* (40% of the comments), *Traffic* (34%) and *Personal safety concern* (29%). Figure 3 shows all the subjects selected, as well as their occurrence in the same data entry (edges) and individual frequencies (node sizes) within the dataset. The edge colour represents the occurrence rate of responses containing two node-subjects. The graph shows a clear cluster around *personal safety*, linking together *walking* and *cycling*. Moreover, these two subjects are often selected with *traffic* and *traffic speed* within the same responses, as shown by the darker coloured edges. This is not surprising since both modes of travel commonly occur within the realm of traffic, and they are affected by its dynamics. The frequency of such co-occurrence in the comments highlights the importance of well-designed infrastructure to ensure the safe coexistence of pedestrians, cyclists, and vehicles.



■ **Figure 4** Word cloud: language patterns in the EBLN free-text comments.

3.3 Language and subjects patterns based on sentiments

In the final part of our analysis, we investigated patterns between the response subjects, the language used in the comments and the respondents' feelings. Given the complexity of the subjects' structure, we decided to group all the subjects into five main categories, following the naming convention used by Bristol City Council:

- **Environment:** air quality, traffic noise, street environment, trees and greenery on street, street trees and planting, littering and bins;
- **Safety:** personal safety concern, traffic speeds, traffic, street lighting;
- **Community:** play space for children, seating and benches;
- **Infrastructure:** footways, crossing, cycle lanes, cycle parking, cycle hangar, bike hangar, bus stop accessibility, EV charging point, car club bays, parking restrictions, blue badge parking;
- **Travel:** walking, cycling, bus services, e-scooters, e-bikes.

We analysed the sentiment distribution within the above categories. Figure 2b shows a notable presence of negative statements in the safety and environment categories, while the community category displays a more balanced distribution of sentiments. This aligns with the observations from the word clouds in Figure 4. In the word clouds, we included the *negative* and *mostly negative* labelled responses in the negative group, and the *positive* and *mostly positive* responses in the positive one, while excluding neutral comments. This approach allows us to accentuate the contrast between the words used to express positive and negative opinions. The analysis of the free-text comments in the EBLN data revealed a predominance of negative terms associated with the environment and travel aspects. Conversely, positive words were more linked to community and green areas. It is worth noticing, the word *road* in the negative cluster, and *street* in the positive one. This distinction reflects the perception that roads typically denote larger, traffic-intensive settings, while streets refer to smaller-scale entities. The sentiment contrast between these terms highlights how respondents express their experiences in relation to urban spaces and transportation infrastructure. Finally, we observe that specific road and area names in the word cloud, such as Beaufort Road, Church Road (negative), and Troopers Hill (positive), correspond to the clustering of negative and positive pins on Figure 1b. We can therefore infer that respondents perceive these locations as areas in need of improvement or additional attention.

4 Conclusions and Future Work

In this paper, we have analysed East Bristol Liveable Neighbourhood (EBLN) online review data to understand the aspects of neighbourhoods perceived by people and identify potential problems. EBLN is a trial project and the dataset used in this study comprises 540 geo-located



contributions. By analysing this dataset, we found that the majority of the respondents reside within the survey area. They tended to emphasize the negative aspects of their neighbourhood, and identify the names of areas and roads associated with positive and negative sentiments in their comments. We also found that most of the negative comments were linked to main roads and junctions while most of the positive comments were linked to community and green areas. The results of this study provide promising preliminary evidence for urban planning and decision-making. Our analysis approach can be applied to a full-scale project. Given the emergent use of public neighbourhood reviews in recent years, it can also be used for similar projects conducted by other cities such as Glasgow and Bath.



There are a number of directions for future work. First, we can extend our analysis to include the reasons and suggestions given by the contributors. Second, evaluate and apply AI-based Natural Language Processing (NLP) techniques for sentiment and semantic analysis of the free-text comments. This would help urban planners to analyse and deduce sentiment and topics from free-text surveys.



References

- 1 Greg Brown, Pat Reed, and Christopher M Raymond. Mapping place values: 10 lessons from two decades of public participation gis empirical research. *Applied Geography*, 116:102156, 2020.
- 2 Greg Brown and Delene Weber. Public participation gis: A new method for national park planning. *Landscape and urban planning*, 102(1):1–15, 2011.
- 3 Greg Brown, Delene Weber, and Kelly de Bie. Is ppgis good enough? an empirical evaluation of the quality of ppgis crowd-sourced spatial data for conservation planning. *Land use policy*, 43:228–238, 2015.
- 4 Geisa Bugs, Carlos Granell, Oscar Fonts, Joaquín Huerta, and Marco Painho. An assessment of public participation gis and web 2.0 technologies in urban planning practice in canela, brazil. *Cities*, 27(3):172–181, 2010.
- 5 Nehal Mahmoud Elmahdy, RR Kamel, and Rania Nasreldin. Contextualizing urban liveability indicators to create liveable neighbourhoods. *International Journal of Engineering Research and Technology*, 14(1):56–68, 2021.
- 6 Yingjie Hu, Chengbin Deng, and Zhou Zhou. A semantic and sentiment analysis on on-line neighborhood reviews for understanding the perceptions of people toward their living environments. *Annals of the American Association of Geographers*, 109(4):1052–1073, 2019.
- 7 Samaneh Khaef and Esfandiar Zebardast. Assessing quality of life dimensions in deteriorated inner areas: A case from javadieh neighborhood in tehran metropolis. *Social indicators research*, 127:761–775, 2016.
- 8 Guy Lansley and Paul A Longley. The geography of twitter topics in london. *Computers, Environment and Urban Systems*, 58:85–96, 2016.
- 9 Jasmine Lau Leby and Ahmad Hariza Hashim. Liveability dimensions and attributes: Their relative importance in the eyes of neighbourhood residents. *Journal of construction in developing countries*, 15(1):67–91, 2010.
- 10 Xiaojun Liu and Wei Hu. Attention and sentiment of chinese public toward green buildings based on sina weibo. *Sustainable cities and society*, 44:550–558, 2019.
- 11 Bernd Resch, Anja Summa, Peter Zeile, and Michael Strube. Citizen-centric urban planning through extracting emotion information from twitter in an interdisciplinary space-time-linguistics algorithm. *Urban Planning*, 1(2):114–127, 2016.
- 12 Zhifang Wang, Zhongwei Zhu, Min Xu, and Salman Qureshi. Fine-grained assessment of greenspace satisfaction at regional scale using content analysis of social media and machine learning. *Science of the Total Environment*, 776:145908, 2021.

Building Alternative Indices of Socioeconomic Status for Population Modeling in Data-Sparse Contexts

Angela R. Cunningham  
Oak Ridge National Laboratory, TN, USA

Joseph V. Tuccillo  
Oak Ridge National Laboratory, TN, USA

Tyler J. Frazier  
Oak Ridge National Laboratory, TN, USA

Abstract

Population modeling requires clear definitions of socioeconomic status (SES) to ensure overall estimate accuracy and locate potentially underserved subpopulations. This presents a challenge as SES can be measured in myriad ways and for divergent purposes, and the data required to calculate these metrics may be lacking, particularly in low and middle income countries (LMICs). To support more refined SES measurement, we explore improvements upon the Demographic and Health Survey's (DHS) Wealth Index (DHS-WI) using alternative characterizations of SES based on multiple correspondence analysis (MCA) and hierarchical clustering. We produce the MCA-derived metrics first on a full suite of household economic, demographic, and dwelling variables, then on a reduced set of occupant-only SES characteristics. We explore the utility of these metrics relative to DHS-WI based on their ability to 1) differentiate DHS household types and 2) identify mixtures of SES levels within DHS samples and mapped at high spatial resolution. We find that our full suite MCA yields more clearly defined SES segments and that our reduced MCA delineates occupant SES most clearly, suggesting potential pathways to improve upon the DHS-WI in future population modeling efforts for LMICs.

2012 ACM Subject Classification Information systems → Geographic information systems

Keywords and phrases Demographic and Health Survey, multiple correspondence analysis, population modeling, socioeconomic status, spatial statistics

Digital Object Identifier 10.4230/LIPIcs.GIScience.2023.25

Category Short Paper

Funding This material is based upon the work supported by the U.S. Department of Energy under contract no. DE-AC05-00OR22725.

Acknowledgements Our thanks to Daniel Adams and Clinton Stipek. This manuscript has been authored by UT-Battelle, LLC under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes. The Department of Energy will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (<http://energy.gov/downloads/doe-public-access-plan>).



© Angela R. Cunningham, Joseph V. Tuccillo, and Tyler J. Frazier;
licensed under Creative Commons License CC-BY 4.0

12th International Conference on Geographic Information Science (GIScience 2023).

Editors: Roger Beecham, Jed A. Long, Dianna Smith, Qunshan Zhao, and Sarah Wise; Article No. 25; pp. 25:1–25:7

Leibniz International Proceedings in Informatics



LIPIC Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Introduction

To improve our ability to locate and address population-related challenges like climate change resiliency and disaster response, we need accurate and precise population estimates that attend to socioeconomic status (SES). SES – variously defined – is predictive of the density at which a population lives [15], as well as life expectancy, [8], lifetime mobility [2], and consumption patterns [9]. However, building appropriate SES definitions is challenging. Measures of SES variably incorporate income, accumulated wealth, education, occupation, cultural markers, demographic factors, resource/infrastructure access, national policies, or embeddedness within the international economy, and are shaped by data availability, and research and policy goals [2][9].

In this paper, we describe efforts to delineate and map SES groups from data recorded in Ghana’s 2014 Demographic and Health Survey (DHS) [5]. Ghana’s survey, like all DHS surveys, is a nationally representative sample collected by a local statistical service in concert with the United States Agency for International Development (USAID), largely for the purposes of monitoring child and maternal health in LMICs. The DHS provides its own Wealth Index (DHS-WI), a measure of household economic status derived from the household’s assets, access to utilities, quality of water sources and toilet facilities, urban/rural status, and the materials with which the household’s dwelling is constructed. DHS reduces these variables to a single metric through a principal components analysis (PCA), taking the first component of the PCA to score households, and dividing the population into quintiles [13]. While DHS-WI is widely used by development agencies and for validation purposes[3], reducing and flattening occupant and dwelling characteristics into a single metric via the first component PCA method complicates the investigation of subpopulation-place relationships whose understanding is central to our own work.

We explore alternative characterizations of SES using DHS data, first based on a full suite of household economic, built environment and demographic variables, and then on a reduced occupant SES only variable set (education and assets). We employ multiple correspondence analysis (MCA) and hierarchical clustering to generate new metrics to compare against DHS-WI, motivated by arguments that 1) PCA is inappropriate for non-continuous variables and 2) that too much useful information is discarded when relying on the first component alone [12][16]. We compare the MCA-derived metrics to the DHS-WI based on distinctiveness and diversity in class labels of respondent households, as well as based on spatial variation in class labels when mapped.

2 Methodology

We began by selecting and preprocessing our variables of interest from Ghana’s 2014 DHS, which consists of 12,831 weighted observations drawn from 30 households randomly sampled from each of the country’s 427 enumeration areas (referred to by DHS as “clusters”)[6]. We extracted variables pertaining to the built environment (source of drinking water; condition of toilet facilities; provision of electricity; number of sleeping rooms; materials used to construct the floor, walls and roof), those that can be used as proxies of wealth (the presence of assets from cars and sewing machines to tables and telephones), the highest level of education in the household, and variables recording basic demographics (household size; age and sex of household head). Following [14], we aggregated building material variables into natural, rudimentary, and finished categories, water and toilet facilities into three quality grades, and binned quantitative variables. We implemented basic data cleaning tasks such as the imputation of missing data using methods specialized for MCA [7].

We conducted our MCAs on the full suite of demographic, occupant SES (assets/education), and aggregated dwelling variables and on a reduced set of only occupant SES variables using R's `FactoMineR` package [7]. We then segmented the individual (household) mappings from each MCA run using `FactoMineR`'s hierarchical clustering method. We set the number of components used by the clustering function to the number of dimensions required to capture over 75% of the variance, thus the full suite and reduced MCAs used seven and four dimensions, respectively. Instead of identifying the number of segments a priori, we allowed the function to select an appropriate number of segments based on minimizing within-cluster inertia for each partition, resulting in three household segments for each MCA run. We also derived quintiles from the first component of our MCAs for the sake of comparison with the DHS-WI.

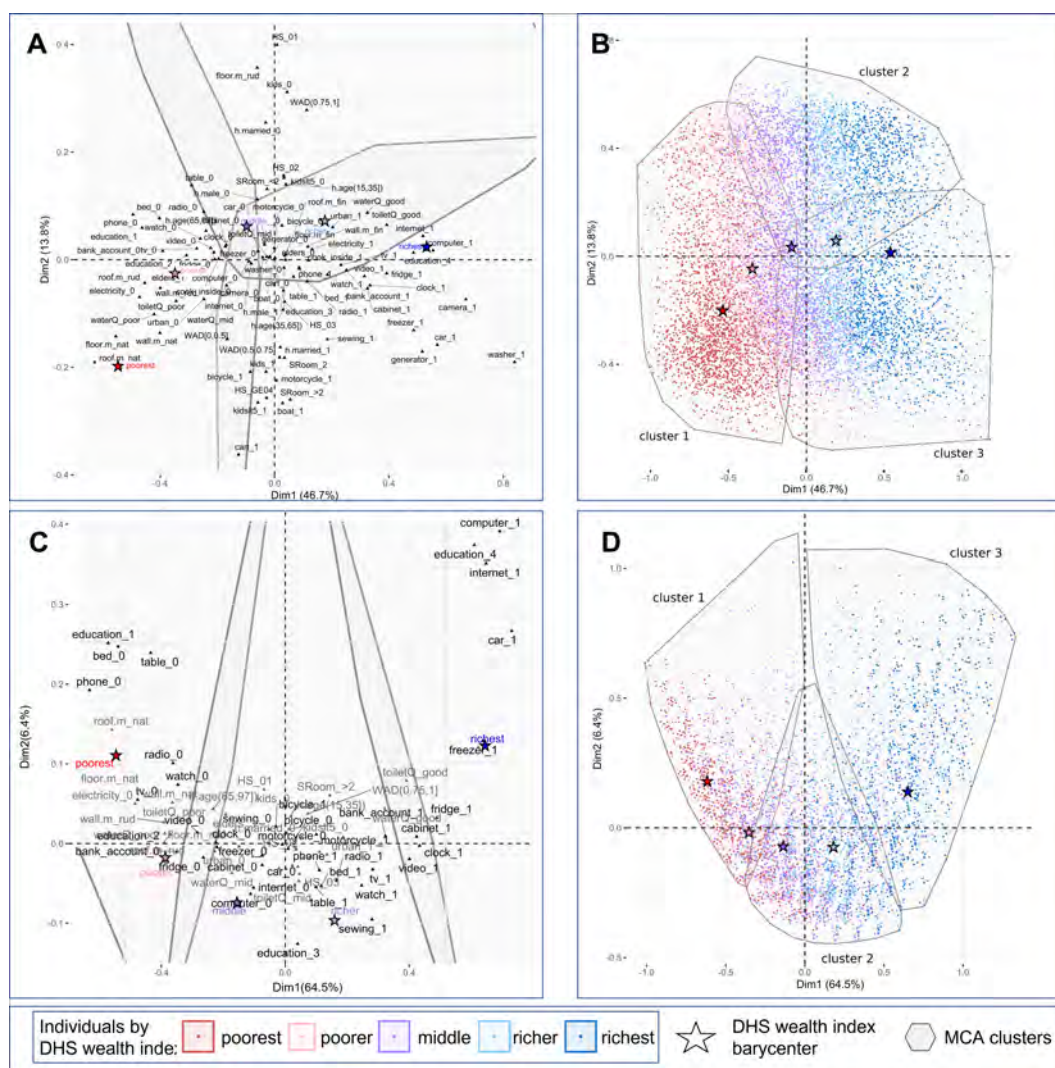
We evaluated the MCA-derived labels against the DHS-WI based on 1) distinctness of the household types they encompass and 2) diversity in household types by DHS sampling cluster. To measure distinctness of households by segment, we calculated silhouette scores, averaged across all variables, built environment variables, occupant SES variables, and demographic variables. Very compact, well separated clusters would score 1; completely overlapped clusters would score -1.

To measure diversity of household types within DHS sampling clusters we calculated Shannon equitability scores, using R's `vegan` package [11] to calculate Shannon entropy and normalizing these figures by the natural log of the number of classes. A score of 0 indicates complete dominance of one label within a sampling cluster; 1, that all labels are present in equal measure. Taken together, these distinctness and diversity measures enable comparison between the DHS-WI and MCA-derived metrics based on both survey-specific and geographic properties of the DHS.

Finally, we compared spatial variation between the DHS-WI and the MCA-based metrics. For each metric, class prevalences (proportions) were described as planar point patterns (PPPs) modeled from 30 random displacements of the DHS sampling cluster centroids (GPS points) for Ghana (within 2km, and 5km-10km displacement zones for urban and rural areas, respectively) [1]. We then apply an edge corrected, absolute risk function to each PPP at 500m resolution to estimate gridded class prevalences that average as final estimates [4]. In future work, these estimates could be combined with gridded population totals to estimate counts of households by SES class; for now they simply describe the expected mixture of groups.

3 Results

Figure 1 visualizes the MCA-derived segments (full suite: panels A, B; reduced: panels C, D) (light gray hulls) relative to projected DHS-WI labels (red to blue: poorest to richest), based on active MCA variables (black text) and descriptive supplemental variables (gray text). For both MCA-derived metrics, dwelling and occupant variables expected to be associated with lower SES (natural or rudimentary building materials, poor quality water and toilet facilities; low education, lack of assets) fall to the left side of the factor maps (panels A and C), while higher household education and increased assets and dwelling quality variables fall to the right. As seen in the cluster maps (panels B and D), individuals assigned a DHS-WI from poorest to richest also track left to right. While the reduced MCA's three segments largely align along this left-right axis (64.5% of explained inertia), the full suite MCA's segments also split across the second dimension which appears to be influenced by demographics: household size (HS), having children (kids), head's marital status (h.married), and working age adult



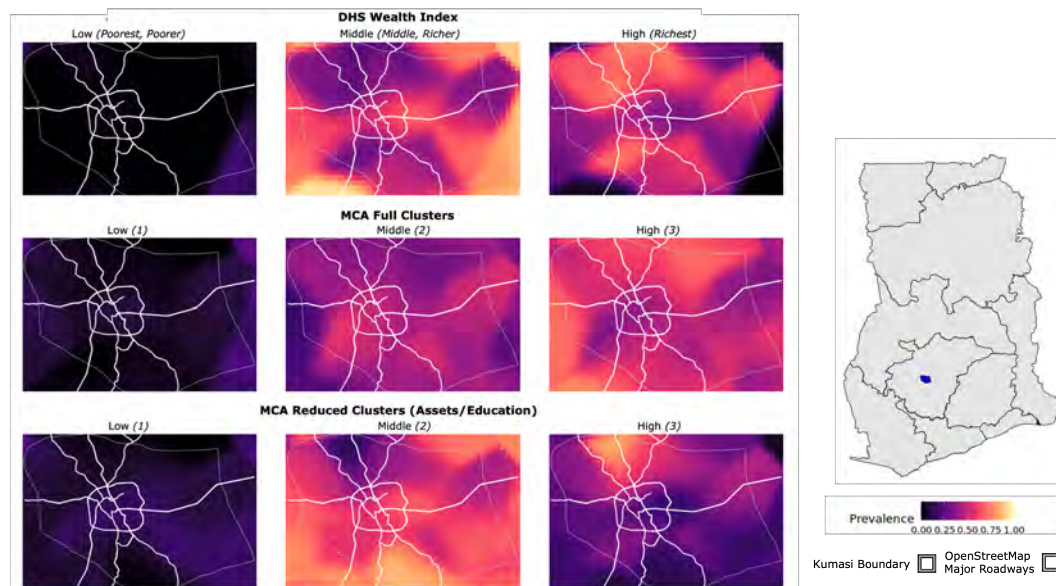
■ **Figure 1** Visualizing segmentation results. A: factor map of full-suite MCA. B: cluster map of full suite MCA. C: factor map of occupant SES MCA. D: cluster map of occupant SES MCA.

proportion (WAD). Given that scholarly opinion is divided as to how such demographic factors should be used in the calculation of SES, our results invite further investigation of this dynamic [8][13].

Table 1 compares the DHS-WI and MCA-derived metrics (both quintiles and segments) based on diversity and distinctness criteria. In terms of **household type distinctness**, the MCA-derived metrics generally improve upon the DHS-WI: full suite MCA segments more cleanly distinguish household types than the DHS-WI when measured across all variables or the variable subsets, while the reduced MCA segments yield the highest silhouette score for the occupant SES variables (though lower ones for built environment variables). Comparing the metrics based on **diversity**, reduced occupant SES segments tend to be considerably more mixed within DHS sampling clusters than the DHS-WI or full suite MCA. Comparing the quintile and segment-based MCA metrics, we also note that allowing for the expression of additional MCA dimensions (via hierarchical clustering) generally improves distinctness/diversity over reliance purely on the first MCA component.

■ **Table 1** Distinctiveness and diversity measures. Silhouettes measured from -1 (complete overlap) to 1 (complete separation), diversity from 0 (one class present) to 1 (all classes equally present).

	DHS-WI	Full suite MCA quintile	Full suite MCA segments	Reduced MCA quintile	Reduced MCA segments
Household Distinctness (Mean Silhouette Width)					
All variables	0.0364	0.0635	0.1882	0.0639	0.1231
Built environment	-0.0163	-0.0075	0.0501	-0.0601	-0.0266
Occupant SES	0.0044	0.0567	0.0988	0.0953	0.2047
Demographic	-0.0912	-0.0430	0.1579	-0.0384	-0.0319
DHS Cluster Diversity (Mean Shannon Equitability)	0.5512	0.7037	0.4995	0.7895	0.7082



■ **Figure 2** Prevalence estimates of SES classes for the DHS-WI, full suite MCA segments, and reduced occupant SES only MCA segments in Kumasi, Ghana.

Figure 2 displays spatial variation in the prevalence estimates of SES segments by metric for Kumasi, Ghana’s second largest city. Consistent with the diversity measures in Table 1, the MCA-based metrics reveal increased mixing of SES classes. In particular, the reduced MCA segmentation shows an increased presence of households associated with low SES (segment 1) compared to the DHS-WI, which features virtually no households with “poorest” and “poorer” labels within Kumasi. Compared to DHS-WI and the full suite MCA, the mapped reduced MCA segments also show increased concentrations of middle-SES households (segment 2) in the south-central area, and high-SES households (segment 3) in the north-central area.

4 Concluding discussion

To explore improvements upon the DHS-WI, we developed several alternative SES classifications based on MCA. Comparing the DHS-WI to the MCA-derived metrics reveals that the former may be both 1) too broad in its scope (subjects/variables) and 2) too reductionist

(one-dimensional) to clearly delineate household SES. Paring our metrics down from an all-inclusive MCA (full suite) to assets and education alone (reduced), we remained able to identify SES categories in alignment with an intuitive socioeconomic gradient. This suggests that in future efforts we can eschew mixing subsets of variables (i.e. built environment with occupant SES) whose relationships we would prefer to explicitly measure. Models leveraging such subpopulation-place relationships could potentially be applied in data sparse contexts to predict occupant SES where this information has not been observed but built environment characteristics have. Further, as demonstrated in Figure 2, these MCA-based metrics reveal more nuanced spatial patterns than the DHS-WI, including potential residential locations of urban poor populations, which in turn can lend detail to characterizations of the built environment extracted from remotely-sensed imagery [10]. Work remains in data engineering, testing modeling alternatives, and external validation, yet this work takes an important step in advancing our ability to map and model populations accurately and precisely.

References

- 1 Adrian Baddeley, Ege Rubak, and Rolf Turner. *Spatial point patterns: methodology and applications with R*. CRC press, 2015.
- 2 Abhijit V. Banerjee and Esther Duflo. What is middle class about the middle classes around the world? *The Journal of Economic Perspectives: A Journal of the American Economic Association*, 22(2):3–28, 2008. doi:10.1257/jep.22.2.3.
- 3 Guanghua Chi, Han Fang, Sourav Chatterjee, and Joshua E. Blumensack. Microestimates of wealth for all low- and middle-income countries. *Proceedings of the National Academy of Sciences*, 119(3):e2113658119, 2022. doi:10.1073/pnas.2113658119.
- 4 P.J. Diggle. *Statistical analysis of spatial point patterns*. Hodder Education, 2003.
- 5 Ghana Statistical Service, Ghana Health Service, and ICF International. Ghana 2014 Demographic and Health Survey [Dataset]. GHPR72FL.DTA, 2015. URL: <https://dhsprogram.com/pubs/pdf/SR224/SR224.pdf>.
- 6 Ghana Statistical Service GSS, Ghana Health Service GHS, and ICF International. Ghana demographic and health survey 2014. Technical report, Ghana Statistical Service – GSS, Rockville, Maryland, USA, 2015. URL: <http://dhsprogram.com/pubs/pdf/FR307/FR307.pdf>.
- 7 Francois Husson, Julie Josse, Sebastien Le, and Jeremy Mazet. FactoMineR: multivariate exploratory data analysis and data mining, 2022. URL: <https://CRAN.R-project.org/package=FactoMineR>.
- 8 Charles I. Jones and Peter J. Klenow. Beyond GDP? Welfare across countries and time. *American Economic Review*, 106(9):2426–2457, 2016. doi:10.1257/aer.20110236.
- 9 Homi Kharas. The unprecedented expansion of the global middle class: an update. Global Economy and Development Working Paper 100, Global Economy and Development at the Brookings Institution, 2017. URL: <https://www.brookings.edu/research/the-unprecedented-expansion-of-the-global-middle-class-2/>.
- 10 Dalton Lunga, Jacob Arndt, Jonathan Gerrand, and Robert Stewart. Resflow: A remote sensing imagery data-flow for improved model generalization. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:10468–10483, 2021.
- 11 Jari Oksanen, Gavin L. Simpson, F. Guillaume Blanchet, Roeland Kindt, Pierre Legendre, Peter R. Minchin, R. B. O’Hara, Peter Solymos, M. Henry H. Stevens, Eduard Szoecs, Helene Wagner, Matt Barbour, Michael Bedward, Ben Bolker, Daniel Borcard, Gustavo Carvalho, Michael Chirico, Miquel De Caceres, Sebastien Durand, Heloisa Beatriz Antoniazzi Evangelista, Rich FitzJohn, Michael Friendly, Brendan Furneaux, Geoffrey Hannigan, Mark O. Hill, Leo Lahti, Dan McGlenn, Marie-Helene Ouellette, Eduardo Ribeiro Cunha, Tyler Smith, Adrian Stier, Cajo J. F. Ter Braak, and James Weedon. vegan: Community Ecology Package, 2022. URL: <https://cran.r-project.org/web/packages/vegan/index.html>.

- 12 Mathieu J. P. Poirier, Karen A. Grépin, and Michel Grignon. Approaches and alternatives to the Wealth Index to measure socioeconomic status using survey data: a critical interpretive synthesis. *Social Indicators Research*, 148(1):1–46, 2020. doi:10.1007/s11205-019-02187-9.
- 13 Shea O. Rutstein and Kiersten Johnson. The DHS wealth index. DHS Comparative Report DHS Comparative Reports No. 6, ORC Macro, Calverton, Maryland, 2004. URL: <https://www.dhsprogram.com/publications/publication-cr6-comparative-reports.cfm>.
- 14 Jeroen Smits and Roel Steendijk. The International Wealth Index (IWI). *Social Indicators Research*, 122:65–85, 2015. doi:10.1007/s11205-014-0683-x.
- 15 Dana R. Thomson, Forrest R. Stevens, Robert Chen, Gregory Yetman, Alessandro Sorichetta, and Andrea E. Gaughan. Improving the accuracy of gridded population estimates in cities and slums to monitor SDG 11: Evidence from a simulation study in Namibia. *Land Use Policy*, 123, 2022. doi:10.1016/j.landusepol.2022.106392.
- 16 Pierre Traissac and Yves Martin-Prevel. Alternatives to principal components analysis to derive asset-based indices to measure socio-economic position in low- and middle-income countries: the case for multiple correspondence analysis. *International Journal of Epidemiology*, 41(4):1207–1208, 2012. doi:10.1093/ije/dys122.

Uncertainty in Causal Neighborhood Effects: A Multi-Agent Simulation Approach

Cécile de Bézenac  

University of Leeds, UK

The Alan Turing Institute, London, UK

Abstract

Interaction between individuals within an environment can result in complex patterns that a statistical analysis is unable to disentangle. The resulting social structure may pose important challenges for the identification of causal relations between variables using only observational data. In particular, the estimation of contextual or neighborhood effects will depend on the spatial configuration under study and the morphology of the areas used to define them. The relevant interpretation of estimates is hence put into question. I suggest adopting a Agent Based Modeling (ABM) approach to study the uncertainty of neighborhood effect estimations within complex spatial systems. An Approximate Bayesian Computing algorithm is used to quantify the uncertainty on the underlying processes that may lead to such estimations. An ABM model of spatial segregation is implemented to illustrate this method.

2012 ACM Subject Classification Human-centered computing

Keywords and phrases Spatial causal inference, neighborhood effects, uncertainty, Agent Based Modeling, Pattern Oriented Modeling

Digital Object Identifier 10.4230/LIPIcs.GIScience.2023.26

Category Short Paper

1 Introduction

The endeavour to generate causal rather than associational claims requires identifying the sufficient conditions for real effects to be estimated. However, their identification can prove to be very challenging in the presence of social and spatial complexity. Namely, interaction between individuals within a geographic environment can result in spatial patterns that a statistical analysis is unable to disentangle. Causal studies that consider factors of a spatial nature however rarely question the nature of space or the different ways in which it may transform the causal analysis [10].

In this paper I consider a category of spatial exposures, the neighborhood effects: the causal effect on an outcome of living in a given area versus living elsewhere. The underlying questions that such exposures raise concern the implications of spatial assumptions on the causal nature of estimated effects. In particular, how does the way we assign neighborhoods to a delimited area impact our causal claims? This question interrogates the inherent uncertainty linked to space in worlds where individuals are in constant movement, in interaction with their peers and with their surroundings. I present some of the challenges linked to the estimation of area-level effects in observational studies. I propose an Agent Based modeling (ABM) approach to generate various forms of spatial complexity against which statistical models may be tested. The exploration of spatial configurations represents an interesting starting point to analyse the so-called neighborhood effects. In particular, ABM exploration methods such as Approximate Bayesian Computing (ABC) offer the means to evaluate the relevance of spatial estimands given complexity assumptions. An illustration of this approach is presented using a simple Schelling model of segregation.



© Cécile de Bézenac;

licensed under Creative Commons License CC-BY 4.0

12th International Conference on Geographic Information Science (GIScience 2023).

Editors: Roger Beecham, Jed A. Long, Dianna Smith, Qunshan Zhao, and Sarah Wise; Article No. 26; pp. 26:1–26:6

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

2 Challenges of estimating and interpreting neighborhood effects

2.1 What are neighborhood effects?

In the literature, one commonly refers to neighborhood effects as the independent effect of a neighborhood on one or multiple outcome [8, 4]¹. For instance, epidemiology studies may compare health outcomes in “poor” versus “rich” areas [17]. Behind this definition lies a number of strong assumptions: what spatial attributes (shape and scale) best describe the neighborhood given the research question? Through what mechanisms can this geometric object be thought to have an independent effect on individual outcome? Finally what are the conditions for these area-level effects to be interpreted causally?

Multiple causal pathways may generate a dependence between the spatial configuration and the individual outcome of interest [4]. The social structure can be thought to play an important role in the shaping of place and the creation of an area identity. Furthermore, social influence or contagion mechanisms may exacerbate or spread exposure effects within an area. Other forms of spatial processes may impact individuals locally such as pollution, crime or the presence of green spaces, etc. The combination of these spatial and social processes can generate important spatial heterogeneity that is typically approximated by variability between specific neighborhood attributes.

2.2 Challenges for estimation

Methods have been developed to estimate neighborhood effects and account for spatial heterogeneity [3]. Namely, multi-level regression models assume some spatial hierarchical structure and allow for heterogeneity by including both fixed and random area-level effects [8]. Still, the interpretation of results are exposed to serious challenges.

The identification of such effects is threatened for one by spatial confounding and at times, complete confounding. A selection process may render the comparison of observations in different areas impossible [1]. This process is further enhanced by social interaction and the non-independence of observations. In a causal analysis, this is referred to as interference or spill-over effects [13].

Finally, spatial and social phenomena can rarely be confined within fixed, arbitrary geographic borders. The choice of spatial areas can lead to the misspecification of neighborhoods that may not reflect any empirical reality. This geographic problem is known as the Modifiable Areal Unit Problem (MAUP) [6].

The previous challenges suggest that the primary danger with mapping the social onto fixed spatial boundaries does not pertain so much to the approximation as it does to the causal interpretation of these so-called neighborhood effects. While accounting for the latter has proven to be insightful for the study of health outcome, employment or education [16, 9], very little work has specifically considered the uncertainty introduced by spatial assumptions. The relation between the misspecification of spatial properties and the bias in results should be looked into. The potential of ABM to generate and analyse the uncertainty surrounding estimates is presented below.

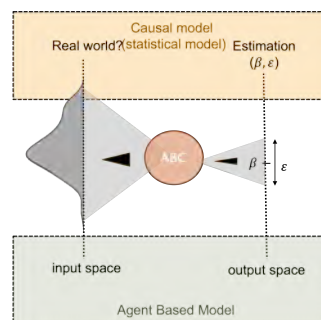
¹ In this paper, we interchangeably use the terms area effects and neighborhood effects when considering small scale spatial configurations

3 A multi-simulation approach

In order to generate artificial complex systems from which granular data may be extracted, we adopt an Agent Based Model (ABM) approach. These models are particularly adapted to create emergent properties from the bottom up by allowing the modeler to build heterogeneous interacting agents while maintaining full control over the micro-level process [12]. The output of these ABM can be placed under the microscope of the same statistical models typically used in observational studies, of which: models that include spatial neighborhood effects.

Many validation methods have been developed to evaluate the performance of ABM for modelling real world systems. Some of these are able to integrate both empirical information and some level of uncertainty on the underlying process. One such method is Approximate Bayesian Computing that approaches a posterior distribution for the parameters of the ABM through typical MCMC algorithms [7]. The general idea is that given information on the system one is modeling, a distribution over the parameter space can be proposed to reflect the probable worlds that may have resulted in similar observations. Samples of the input space are drawn and either selected or rejected according to a proximity criteria, usually determined by an error threshold ϵ . This threshold represents a level of uncertainty in the output space: how precise is the information on the real system? A more detailed description of ABC can be found in [14].

I suggest using ABC, not to evaluate the ABM but to interrogate the relevance of spatial assumptions in analytical models. The questions that this framework should be able to answer are: How biased are the estimations of neighborhood effects when obtained from data generated by the ABM? What other complex systems and mechanisms may be considered as possible generators of these estimations, given the statistical confidence levels. The notion of *sufficiency* of the statistical approach is introduced here as its performance in distinguishing between multiple causal mechanisms can be analysed. The uncertainty considered here links the *equifinality* of complex systems [15] to the epistemic uncertainty of classical models [2]. In the following, I present a simple illustration of the use of this framework on a spatial segregation model, the Schelling model. This model was chosen as an illustration for its simplicity and its focus on spatial interaction between agents.



■ **Figure 1** Schema of the ABM-ABC framework. Here β represents the objective/estimation and ϵ the threshold/standard error.

4 Illustration and results

4.1 The Schelling model

The Schelling model [11] can be described in the following way: on a square lattice a number of blue (exposed) and red pawns (non-exposed) occupy individual cells. These pieces will move to an empty cell if the percentage of a piece’s same-color neighbors (Moore neighborhood) is lower than a predetermined threshold H (designating the homophily level). At each step, the elements on the board will be displaced according to their surrounding composition until they can no longer move or until all pieces are satisfied. The dynamics of this system tell a very interesting story as clear segregation patterns emerge without any higher order intervention. The relationship between the agent’s attributes and their environment can be simply translated into a linear form for a given step in time.

$$Y_i(H, C_i, C_{N_i}) = H - C_i - \frac{1}{d_i}g(C_{N_i}) + \frac{2C_i}{d_i}g(C_{N_i}) \quad (1)$$

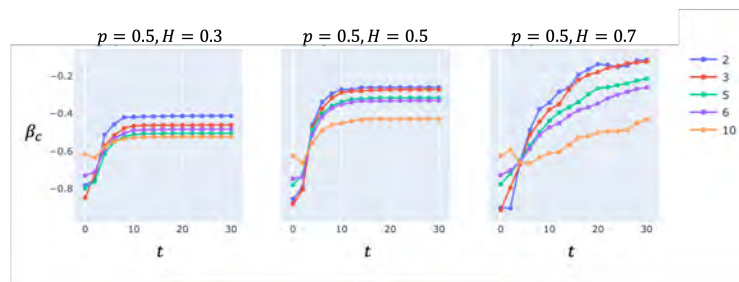
Where Y_i is the outcome for agent i , C_i is their color or the individual “exposure”, $g(C_{N_i})$ is the aggregated prevalence of blue within i ’s ego-neighborhood and finally d_i is size of i ’s neighborhood. Translated into network terms, d_i is the degree of i in the regular graph drawn from the grid structure (for Moore neighborhood, $d_i \leq 8$). Note that there is clear violation of the no-interference assumption as agents outcome will depend on their own color and the color of peers. I consider the simple data scenario where the graph (or ego) neighborhood of pieces is not known but interaction is assumed constrained to fixed predetermined areas. Neighborhood exposure is then approximated by “blue” prevalence within an area. This specification of neighborhoods is analysed in light of the approximated posterior distribution for two of the models’ key parameters: the homophily level (H) and the probability of agents belonging to the blue group (p). Some of the results obtained for different neighborhood scales are presented below.

4.2 Results

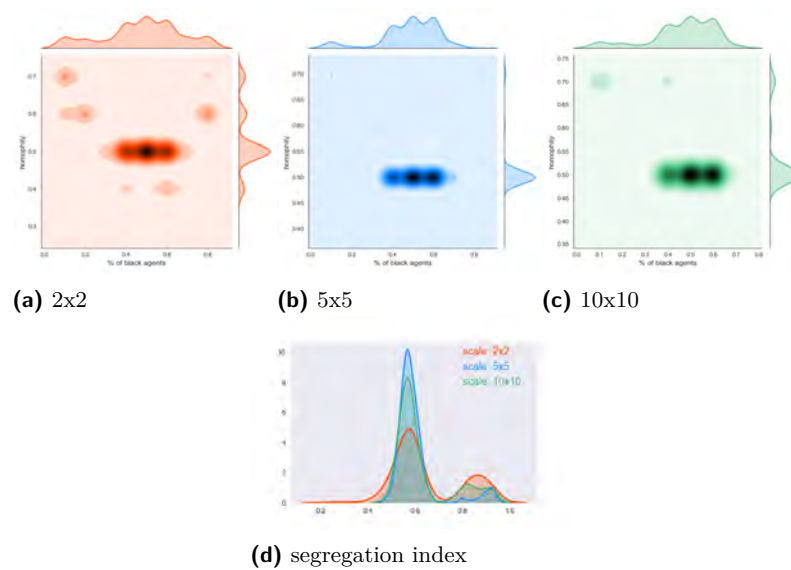
The 40x40 Schelling torus grids are cut up into 4, 25 and 100 different areas, respectively. An ABC is run for each choice of scale:

Samples were drawn from a uniform prior over the input space. A linear estimation of exposure and neighborhood effects are computed on a random simulation output for $p = 0.5$, $H = 0.5$. The estimation at step $t = 30$ is used as the objective for the ABC algorithm, the standard error of the estimation serves as the restrictive threshold for the rejection process. $N=1000$ particulars are tested against these criteria. Both the ABM model and the ABC algorithm were run using standard Python packages [5].

The estimation of color and neighborhood effects vary with time as the spatial patterns converges to a segregated state. It is quite clear that the choice of scale impacts the quality of the estimation. Larger scales blur the information on micro-level heterogeneity and social borders are hidden within fixed areas. The problem of total confounding may also arise for smaller scales as more and more selective neighborhoods appear. The results of the ABC show that the uncertainty introduced by the spatial approximation does interact with model parameters in a uniform way. While the true homophily level is relatively well identified, a wider range of exposure assignment p may lead to similar estimations. The distribution of the segregation index (as the average similarity of peers) shows that very different spatial patterns can lead to the same interpretation of neighborhood effects: not only are the estimations heavily biased, they do not describe the system sufficiently well. What is interesting to note



■ **Figure 2** Variation in estimation of OLS estimator for color effects (β_c) under misspecification of neighborhood for resp. (2x2), (3x3), (5x5), (6x6), (10x10) area grids.



■ **Figure 3** Posterior distributions for H and p approximated from the ABC and the estimation of color effect for different scales (a-c); Distribution of segregation in the selected simulations for scales 2x2, 5x5 and 10x10 (d).

is the influence of scale on the uncertainty. It appears there exists a scale for which the possibilities are reduced and the posterior distributions are more concentrated (for instance see Fig. 3.b))

5 Discussion

This very simple model was used to illustrate the use of ABM pattern oriented methods to question the uncertainty of causal model estimates. The specific representation of space, here as neighborhood effects, will have an impact on the meaning of the estimands and ultimately, on the appropriate interpretation of estimations. Notions of spatial equifinality in complex systems should be considered to better understand the role of space in social mechanisms. A relevant road-map for spatial causal inference would consider the specific challenges that relate to defining this spatial context. Thinking of causality within a spatial context may imply moving from a paradigm of causal dependence between variables to one of causal mechanisms in a system of spatially embedded agents.

References

- 1 Steven N Durlauf. Neighborhood effects. *Handbook of regional and urban economics*, 4:2173–2242, 2004.
- 2 Scott Ferson and Kari Sentz. Epistemic uncertainty in agent-based modeling. In *7th international workshop on reliable engineering computing*, pages 65–82, 2016.
- 3 A Stewart Fotheringham and Mehak Sachdeva. Modelling spatial processes in quantitative human geography. *Annals of GIS*, 28(1):5–14, 2022.
- 4 George C Galster. The mechanism (s) of neighbourhood effects: Theory, evidence, and policy implications. In *Neighbourhood effects research: New perspectives*, pages 23–56. Springer, 2011.
- 5 Jackie Kazil, David Masad, and Andrew Crooks. Utilizing python for agent-based modeling: The mesa framework. In Robert Thomson, Halil Bisgin, Christopher Dancy, Ayaz Hyder, and Muhammad Hussain, editors, *Social, Cultural, and Behavioral Modeling*, pages 308–317, Cham, 2020. Springer International Publishing.
- 6 David Manley, Robin Flowerdew, and David Steel. Scales, levels and processes: Studying spatial patterns of british census variables. *Computers, environment and urban systems*, 30(2):143–160, 2006.
- 7 Josie McCulloch, Jiaqi Ge, Jonathan A Ward, Alison Heppenstall, J Gareth Polhill, and Nick Malleson. Calibrating agent-based models using uncertainty quantification methods. *Journal of Artificial Societies and Social Simulation*, 25(2), 2022.
- 8 J Michael Oakes. The (mis) estimation of neighborhood effects: causal inference for a practicable social epidemiology. *Social science & medicine*, 58(10):1929–1952, 2004.
- 9 Francis A Pearman. Gentrification and academic achievement: A review of recent research. *Review of Educational Research*, 89(1):125–165, 2019.
- 10 Brian J Reich, Shu Yang, Yawen Guan, Andrew B Giffin, Matthew J Miller, and Ana Rappold. A review of spatial causal inference methods for environmental and epidemiological applications. *International Statistical Review*, 89(3):605–634, 2021.
- 11 Thomas C Schelling. Dynamic models of segregation. *Journal of mathematical sociology*, 1(2):143–186, 1971.
- 12 Flaminio Squazzoni. The impact of agent-based models in the social sciences after 15 years of incursion. *The Impact of Agent-Based Models in the Social Sciences after 15 Years of Incursion*, pages 1000–1037, 2010.
- 13 Eric J Tchetgen Tchetgen and Tyler J VanderWeele. On causal inference in the presence of interference. *Statistical methods in medical research*, 21(1):55–75, 2012.
- 14 Brandon M Turner and Trisha Van Zandt. A tutorial on approximate bayesian computation. *Journal of Mathematical Psychology*, 56(2):69–85, 2012.
- 15 Konstantina Valogianni and Balaji Padmanabhan. Causal abms: Learning plausible causal models using agent-based modeling. In *The KDD’22 Workshop on Causal Discovery*, pages 3–29. PMLR, 2022.
- 16 Maarten Van Ham, David Manley, Nick Bailey, Ludi Simpson, and Duncan Maclennan. Neighbourhood effects research: New perspectives. In *Neighbourhood effects research: New perspectives*, pages 1–21. Springer, 2011.
- 17 Blair Wheaton, Rosane Nisenbaum, Richard H Glazier, James R Dunn, Catharine Chambers, et al. The neighbourhood effects on health and well-being (nehw) study. *Health & place*, 31:65–74, 2015.

Uncovering Spatiotemporal Patterns of Travel Flows Under Extreme Weather Events by Tensor Decomposition

Zhicheng Deng ✉ 

School of Urban Planning and Design, Peking University Shenzhen Graduate School, Shenzhen, Guangdong, China

Key Laboratory of Earth Surface System and Human-Earth Relations of Ministry of Natural Resources of China, Peking University Shenzhen Graduate School, Shenzhen, Guangdong, China

Zhaoya Gong¹ ✉ 

School of Urban Planning and Design, Peking University Shenzhen Graduate School, Shenzhen, Guangdong, China

Key Laboratory of Earth Surface System and Human-Earth Relations of Ministry of Natural Resources of China, Peking University Shenzhen Graduate School, Shenzhen, Guangdong, China

Pengjun Zhao ✉

School of Urban Planning and Design, Peking University Shenzhen Graduate School, Shenzhen, Guangdong, China

Key Laboratory of Earth Surface System and Human-Earth Relations of Ministry of Natural Resources of China, Peking University Shenzhen Graduate School, Shenzhen, Guangdong, China
College of Urban and Environmental Sciences, Peking University, Beijing, China

Abstract

Extreme weather events have caused dramatic damage to human society. Human mobility is one of the important aspects that are impacted significantly by extreme weather. Currently, focus on human mobility research during extreme weather is often limited to the transport infrastructure and emergency management perspectives, lacking a systematic understanding of the spatiotemporal patterns of human travel behavior. In this research, we examine the structural changes in human mobility under the severe rainstorm that occurred on July 20th, 2021 in Zhengzhou, Henan Province, China. Innovatively applying a tensor decomposition approach to analyzing spatiotemporal flows of human movements represented by the mobile phone big data, we extract the characteristic components of human travel behaviors from the spatial and temporal dimensions, which help discover and understand the latent spatiotemporal patterns hidden in human mobility data. This study provides a new methodological perspective and demonstrates that it can be useful for uncovering latent patterns of human mobility and identifying its structural changes during extreme weather events. This is of great importance to a better understanding of the behavioral side of human mobility and its response to external shocks and has significant implications for human-focused policies in urban risk mitigation and emergency response.

2012 ACM Subject Classification Information systems → Geographic information systems

Keywords and phrases Urban travel behavior, Origin-Destination flows, Non-negative CP decomposition, Spatiotemporal analysis

Digital Object Identifier 10.4230/LIPICs.GIScience.2023.27

Category Short Paper

Funding *Zhaoya Gong*: Shenzhen Science and Technology Program JCYJ20220818100810024.

Pengjun Zhao: Shenzhen Science and Technology Program JCYJ20220818100810024.

¹ Corresponding author



© Zhicheng Deng, Zhaoya Gong, and Pengjun Zhao;
licensed under Creative Commons License CC-BY 4.0

12th International Conference on Geographic Information Science (GIScience 2023).

Editors: Roger Beecham, Jed A. Long, Dianna Smith, Qunshan Zhao, and Sarah Wise; Article No. 27; pp. 27:1–27:6

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Introduction

Cities are not simply spaces with idealized morphologies; rather, we should understand them as complex systems composed of networks and flows [1]. Within cities, residents travel by various transportation modes, which is reflected in the spatiotemporal patterns of urban travel behavior, forming different urban rhythms and spatial structures. Understanding these travel behavior patterns is crucial for helping to better understand the complex urban system, thereby bringing implications to people-oriented policies and promoting urban management capabilities.

According to the 2023 IPCC Report on Climate Change [4], human-caused climate change has recently affected the weather and extreme climate in all regions of the world, causing widespread damage and destruction to nature and humanity. Existing research has shown that when faced with extreme weather, the patterns of human mobility can exhibit spatiotemporal characteristics different from those in normal times [10, 3]. However, in the field of GIScience, most studies still adopt the transport infrastructure and emergency management perspectives, utilizing GIS methods to investigate human behavior during disasters or to simulate the evacuation patterns of individuals [9, 5]. These studies lack a systematic understanding of the spatiotemporal patterns of human mobility under external impacts.

With the widespread adoption of smartphones and the development of positioning technology, a vast amount of population activity data has been generated, which is now widely used in urban research [6, 2, 11]. This kind of data contains information about human travel behavior, interactions between different areas of the city, and the spatial structure of the city. The availability of these data and corresponding analysis methods provide possibilities for quantifying and measuring human travel under normal conditions and during natural disasters, as well as analyzing the spatiotemporal patterns of travel flows.

This study utilizes travel flows recorded by mobile phones to construct a tensor of human mobility, and decomposes it using tensor factorization methods to extract the spatiotemporal characteristics. As a case study, this paper focuses on a torrential rain event that occurred on July 20th, 2021 in Zhengzhou, Henan Province, China, which resulted in 380 deaths and affected more than a million people [8]. This paper explores the differences in urban travel behavior under normal conditions and external impacts of the rainstorm, reveals the spatiotemporal patterns hidden in travel behavior, and analyzes the underlying spatiotemporal mechanisms causing changes in human mobility patterns after the rainstorm.

2 Methods and Data

2.1 Methods

Multidimensional travel flows can be organized in a “space-time” format to construct a spatiotemporal tensor. In this study, for the analysis of travel flows, we use Origin-Destination (OD) pairs as the spatial dimension and time as the temporal dimension to store and express the travel flow values between the origins and the destinations.

Tensor decomposition is a low-rank approximation method for tensors. Through tensor decomposition, we can extract the main characteristics and relationships in the travel flows. In this study, the CANDECOMP/PARAFAC (CP) tensor decomposition is chosen, which is a representative and easy-to-understand analysis method, to decompose a high-order tensor into a sum of rank-1 tensors. The principle of decomposition is shown in Figure 1, with the tensor constructed from the original data denoted by $\mathcal{X} \in \mathbb{R}^{I \times J}$. I is the number of OD pairs, J is the length of the time unit, and R is a positive integer, representing the rank in

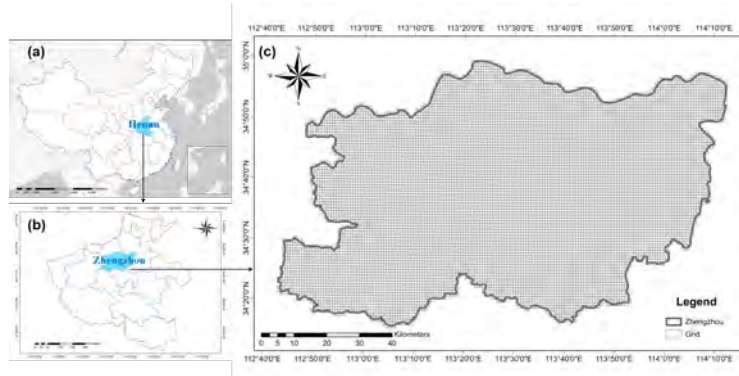
the tensor decomposition, which is the number of features in each dimension. R modes are obtained by the decomposition. $OD_r, T_r (r = 1, \dots, R)$ are the characteristics of the spatial and temporal dimensions, and λ_r is the corresponding weight. Considering the non-negativity of the travel flows, this study uses the Non-negative CP (NNCP) decomposition proposed by Shashua and Hazan [7] to implement.

$$\begin{array}{c}
 \boxed{\mathcal{X} \in \mathbb{R}^{I \times J}} \approx \lambda_1 \times \begin{array}{c} \boxed{T_1 \in \mathbb{R}^{I \times J}} \\ \text{---} \\ \boxed{OD_1 \in \mathbb{R}^{I \times 1}} \end{array} + \lambda_2 \times \begin{array}{c} \boxed{T_2 \in \mathbb{R}^{I \times J}} \\ \text{---} \\ \boxed{OD_2 \in \mathbb{R}^{I \times 1}} \end{array} + \dots + \lambda_R \times \begin{array}{c} \boxed{T_R \in \mathbb{R}^{I \times J}} \\ \text{---} \\ \boxed{OD_R \in \mathbb{R}^{I \times 1}} \end{array}
 \end{array}$$

■ **Figure 1** The CP decomposition for travel flows.

2.2 Data

Zhengzhou is located in the central-northern part of China, and is the capital city of Henan Province. It is also the economic and population center of Henan Province and an important transportation hub for the entire country with an enormous amount of floating population. From July 19th to 23rd, 2021, a torrential rain disaster occurred in Zhengzhou, which broke the historical record of extreme meteorological observation in mainland China, causing heavy casualties and property losses. In this study, we select Zhengzhou as the research area, and divide it into basic spatial units of 1 km * 1 km grids, as shown in Figure 2.



■ **Figure 2** The geographical location of our research area. (a) the research area in China; (b) the research area in Henan Province; (c) the spatial distribution of grids in Zhengzhou.

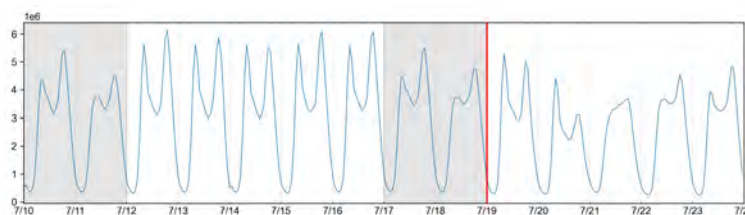
The travel flow data used in this study covers a period of two weeks, from July 10th to July 23rd, 2021, with a time resolution of one hour. In order to remove abnormal values that may affect the experiment, we impose a threshold on the value of OD flows. Only flows with a daily average value greater than 10 are included, resulting in 1,027,568 OD pairs on 336 time units. Moreover, we perform smoothing processing on the experimental data to reduce the effects of anomalies and peaks and obtain results with better reconstruction rates, using a window size of 3, with a convolution kernel set to [0.3, 0.4, 0.3]. The same data quantity are maintained.

Based on the above data, we construct a travel tensor with a size of [1027568, 336]. Each row represents an OD pair, and each column represents an hour. The value in each element represents the flow volume from the origin to the destination within one hour. The first 168 columns represent the first week or normal conditions, and the remaining 168 columns represent the second week or the external impact of Zhengzhou torrential rain disaster.

3 Results

3.1 Original Data Analysis

Firstly, we evaluate and analyze the impact of the torrential rain event on the travel behavior of people based on the original data. The total travel volume during the period from July 10th to 23th is shown in Figure 3. It can be inferred that there exists a daily rhythm in the residents' travel behavior. In addition, due to the impact of the torrential rain in Zhengzhou, the travel volume significantly and relatively decreases after July 19th, which implies a transition from the normal behavior pattern to the abnormal behavior pattern.



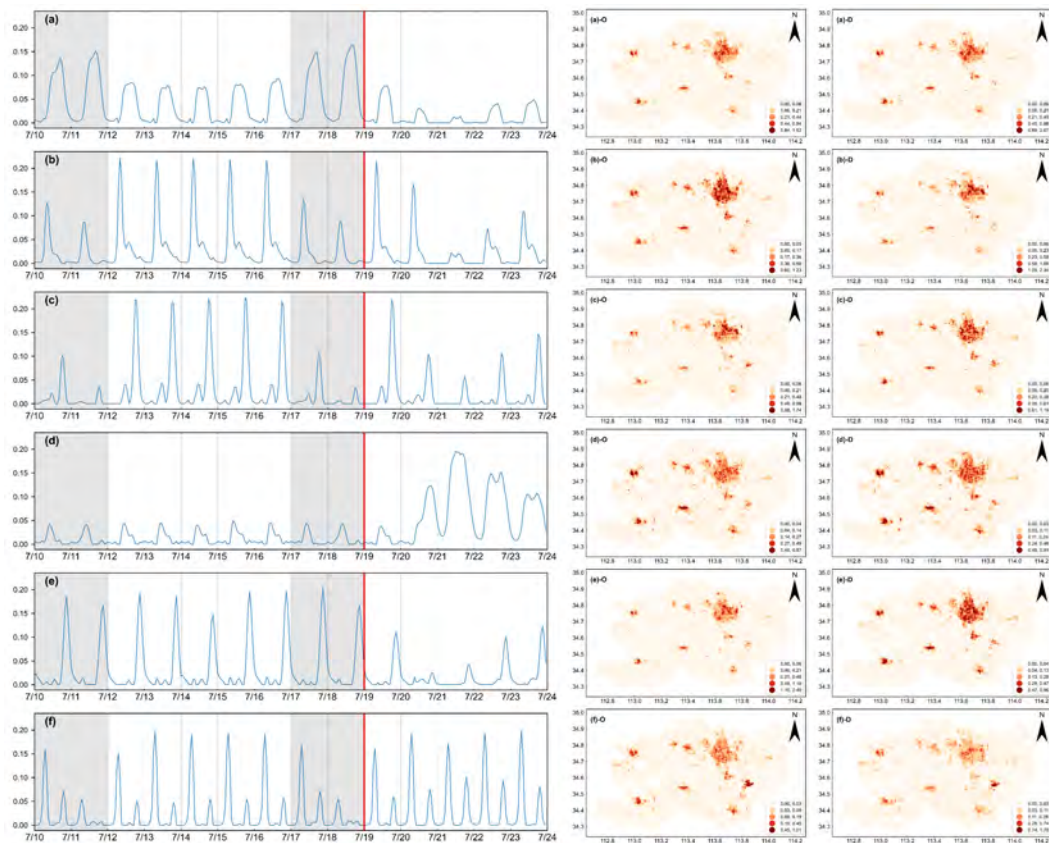
■ **Figure 3** Total travel volume changing with time.

3.2 Decomposition Results

We use NNCP decomposition to extract spatiotemporal modes of travel behavior, obtaining two outputs: temporal patterns and spatial patterns. The rank parameter is selected according to the root mean square error of decomposition results at different ranks. We use the rank 6 at the elbow point as an example for experiment and analysis. Six modes are obtained by decomposition, and their weights are sorted as follows: 87980.86, 73816.80, 61583.89, 59256.50, 57573.24, and 48475.62. The corresponding temporal patterns and spatial patterns are shown in Figure 4. In spatial patterns, values of decomposition results are aggregated according to the origins and destinations.

In the sub-figure of temporal patterns, the gray background represents weekends, and the white background represents weekdays. The red vertical line represents the start day of the torrential rain in Zhengzhou. The most important mode a represents the overall trend of the city's travel behavior. Compared to weekends, weekday travel volumes are lower. On the day before the heavy rain of July 19th, the travel behavior still shows a normal pattern. However, after July 20th, the travel volume decreases greatly and reaches its minimum on July 21st. Mode b can be interpreted as the morning peak mode, which reaches its peak around 7:00–10:00. Compared to mode a, the peak height of mode b is also higher on weekdays than on weekends, and there is a secondary peak in the afternoon (around 14:00). Mode c represents the evening peak mode, and there is a secondary peak at around 12:00, which corresponds to the secondary peak of mode b. Mode d represents the abnormal travel mode caused by external impacts such as torrential rain and other additional information outside the main mode. It increases greatly from July 20th and reaches its peak on July 21st, then drops rapidly. Mode e shows the travel flows after the evening peak during the late night (20:00–23:00), and mode f represents the stronger characteristics of the morning and evening peaks.

We combine the temporal patterns to interpret the spatial patterns. In mode a, the results for the origins and the destinations are relatively similar. The high-value areas mainly include the central city of Zhengzhou, the airport area, and the centers of the county-level



■ **Figure 4** Spatiotemporal patterns of travel flows.

cities. Mode b–O corresponds to modes c–D and e–D. Combined with the temporal patterns, it shows that the origins of the early peak flow are consistent with the destinations of the late peak flow and the late night flow, reflecting the commuting patterns of urban residents for work and life. It can be inferred that the primary residences are Weilai Road Street, Nanyang Road Street, Jingba Road Street, and Tongbai Road Street, etc., while their work locations are Jicheng Road Street (provincial government and other administrative regions), Zhengzhou East Station area, and Zhengzhou Railway Station area. Therefore, it may also contain information about cross-regional travel. Unlike previous modes, mode d reflects the travel patterns during heavy rain periods, with higher values in the areas of county-level city centers, which are more affected by the rainstorm. These areas experience relatively increased travel due to rescue and other activities, while the central city areas have a significant decrease in travel intensity. Mode f is mainly located in the Foxconn Park area. Combined with the time mode, it may reflect the commuting mode of its workers.

4 Conclusions and Discussions

In this study, we use NNCP decomposition to extract and analyze different urban travel patterns before and during the 720 torrential rain event in Zhengzhou, Henan Province, China. We find that there are multiple spatial and temporal patterns. The temporal patterns include morning peak, evening peak, daytime flow, late night flow, and early morning flow. The spatial patterns correspond to the interaction between residence and workplace, and the

interaction between residence and other functional places, and so on. In particular, under the external impact of the torrential rain disaster, people may shift their travel modes to avoid potential risks. Temporally, the travel pattern shows an intense increase from July 20th after the torrential rain, reaching a peak on July 21st, followed by a rapid decline. Spatially, the internal travels within counties are relatively strengthened, and different travel patterns are also observed in the urban area.

This paper innovatively applies a tensor decomposition approach to analyzing spatiotemporal flows of human movements under extreme weather events, effectively extracting different urban travel behavior characteristics under different circumstances, and exploring the response of urban travel behavior patterns to extreme weather. However, there is still room for improvement in this study. Currently, the interpretation of the obtained travel pattern results is based on exploratory analysis and limited to speculative discussions. In the future, more confirmative analysis can be conducted to validate the multi-scale characteristics of travel patterns. For temporal patterns, time series analysis can be used to extract time-frequency domain characteristics, providing descriptions and predictions. For spatial patterns, methods such as network community detection can be used to divide urban areas.

References

- 1 Michael Batty. *The new science of cities*. MIT press, 2013.
- 2 Francesco Calabrese, Mi Diao, Giusy Di Lorenzo, Joseph Ferreira Jr, and Carlo Ratti. Understanding individual mobility patterns from urban sensing data: A mobile phone trace example. *Transportation research part C: emerging technologies*, 26:301–313, 2013.
- 3 Boyeong Hong, Bartosz J Bonczak, Arpit Gupta, and Constantine E Kontokosta. Measuring inequality in community resilience to natural disasters using large-scale mobility data. *Nature communications*, 12(1):1870, 2021.
- 4 IPCC. *Climate Change 2023: Synthesis Report. A Report of the Intergovernmental Panel on Climate Change*. IPCC, Geneva, Switzerland, (in press), 2023.
- 5 Marianna Loli, George Kefalas, Stavros Dafis, Stergios A Mitoulis, and Franziska Schmidt. Bridge-specific flood risk assessment of transport networks using gis and remotely sensed data. *Science of the Total Environment*, 850:157976, 2022.
- 6 Tao Pei, Stanislav Sobolevsky, Carlo Ratti, Shih-Lung Shaw, Ting Li, and Chenghu Zhou. A new insight into land use classification based on aggregated mobile phone data. *International Journal of Geographical Information Science*, 28(9):1988–2007, 2014.
- 7 Amnon Shashua and Tamir Hazan. Non-negative tensor factorization with applications to statistics and computer vision. In *Proceedings of the 22nd international conference on Machine learning*, pages 792–799, 2005.
- 8 The State Council Disaster Investigation Team. Investigation report on '720' heavy rainfall disaster in zhengzhou,henan. Technical report, The State Council of People's Republic of China, 2022.
- 9 Brian Tomaszewski. *Geographic information systems (GIS) for disaster management*. Routledge, 2020.
- 10 Qi Wang and John E Taylor. Patterns and limitations of urban human mobility resilience under the influence of multiple types of natural disaster. *PLoS one*, 11(1):e0147299, 2016.
- 11 Yang Xu, Jiaying Xue, Sangwon Park, and Yang Yue. Towards a multidimensional view of tourist mobility patterns in cities: A mobile phone data perspective. *Computers, Environment and urban systems*, 86:101593, 2021.

GeoQAMap - Geographic Question Answering with Maps Leveraging LLM and Open Knowledge Base

Yu Feng  

Chair of Cartography and Visual Analytics, Technical University of Munich, Germany

Linfang Ding  

Norwegian University of Science and Technology, Trondheim, Norway

Guohui Xiao  

Department of Information Science and Media Studies, University of Bergen, Norway

Abstract

GeoQA (Geographic Question Answering) is an emerging research field in GIScience, aimed at answering geographic questions in natural language. However, developing systems that seamlessly integrate structured geospatial data with unstructured natural language queries remains challenging. Recent advancements in Large Language Models (LLMs) have facilitated the application of natural language processing in various tasks. To achieve this goal, this study introduces GeoQAMap, a system that first translates natural language questions into SPARQL queries, then retrieves geospatial information from Wikidata, and finally generates interactive maps as visual answers. The system exhibits great potential for integration with other geospatial data sources such as OpenStreetMap and CityGML, enabling complicated geographic question answering involving further spatial operations.

2012 ACM Subject Classification Applied computing → Cartography

Keywords and phrases Geographic Question Answering, Large Language Models, SPARQL, Knowledge Base, Wikidata

Digital Object Identifier 10.4230/LIPICs.GIScience.2023.28

Category Short Paper

Funding Dense and Deep Geographic Virtual Knowledge Graphs for Visual Analysis (DFG Grant 500249124).

1 Motivation

The recent progress in Natural Language Processing (NLP), specifically with Large Language Models (LLMs) has demonstrated significant potential for automating a wide range of tasks. The field of GIScience is actively embracing the utilization of artificial intelligence and seeking to enhance traditional workflows through their integration. Within this context, GeoQA (Geographic Question Answering) has emerged as a prominent research area, focusing on the development of intelligent systems capable of answering questions involving geographic entities or concepts. By leveraging the power of NLP and knowledge graph, GeoQA aims to enable more efficient and effective utilization of geographic information for improved decision-making and problem-solving in various domains.

However, geospatial question answering is challenging, primarily because it involves the integration of structured geospatial data with unstructured natural language queries. Geospatial data typically has a structured format that represents spatial relationships, coordinates, and attributes of geographic entities. On the other hand, natural language queries are unstructured and require understanding and interpretation to extract the relevant geospatial information. Current models are mostly based on text or images. ChatGPT is primarily a text-based model and does not have the capability to directly generate maps. The user would only reply with guidance on how to generate a map using conventional software or



© Yu Feng, Linfang Ding, and Guohui Xiao;
licensed under Creative Commons License CC-BY 4.0

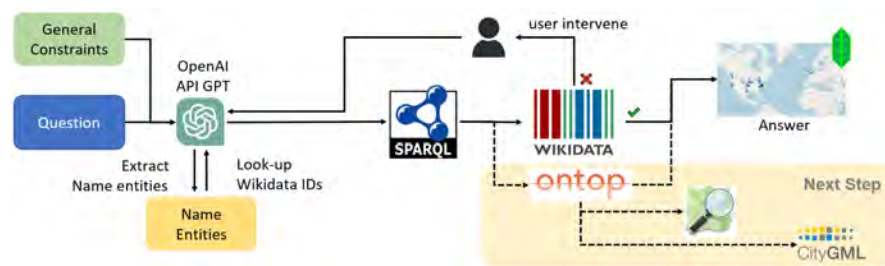
12th International Conference on Geographic Information Science (GIScience 2023).

Editors: Roger Beecham, Jed A. Long, Dianna Smith, Qunshan Zhao, and Sarah Wise; Article No. 28; pp. 28:1–28:7

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany



■ **Figure 1** Overview of the proposed GeoQAMap system.

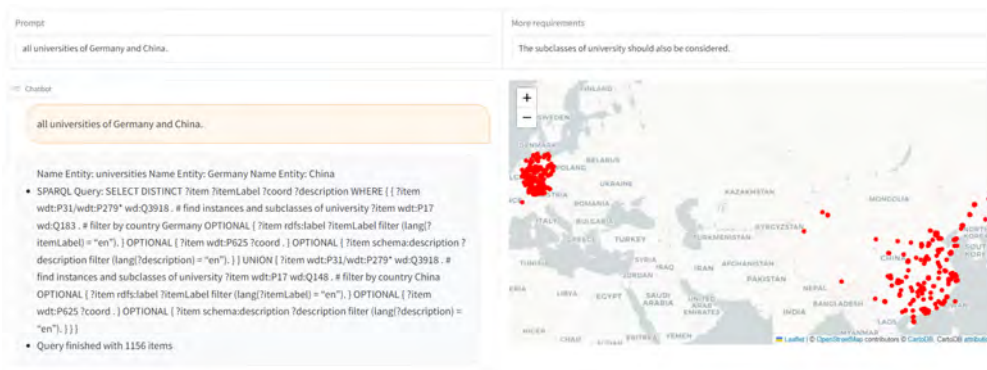
programming languages. On the other hand, there are image-based generative models, such as Midjourney or Stable Diffusion¹, that can generate images containing maps as content. However, it is important to note that these generated map images may not conform to the standard formats and conventions commonly associated with geospatial data [2].

To address this challenge, an intermediary becomes essential to bridge the gap between geospatial data and natural language queries. One potential solution is to utilize SPARQL, a query language specifically designed for querying data stored in the Resource Description Framework (RDF) format. RDF provides a standardized representation for data using subject-predicate-object triples, making it suitable for structured geospatial data. SPARQL is nowadays standard for representation and querying of linked data for semantic web. Furthermore, SPARQL's capabilities have been extended to GeoSPARQL, which incorporates spatial operations, enhancing its utility for handling and analyzing geospatial data.

However, it is worth noting that SPARQL queries often involve complex syntax and rules, making them challenging for end users to grasp and utilize effectively. The intricacies of the language can pose a barrier to entry for individuals who are not familiar with its syntax or who lack technical expertise. To address this challenge, the emerging field of LLMs has provided a promising solution. By leveraging LLMs, natural language queries can be translated into SPARQL queries that can access structured geospatial data stored in RDF format. SPARQL queries can retrieve the relevant information based on the query's spatial constraints, enabling the integration of geospatial data and natural language queries. Since the research leveraging LLM and knowledge system is a rather new research field, there were not yet many applications that demonstrate the ability answering geospatial questions with maps. Only recently, there was one work named *Autonomous GIS* presented by [3]. In their process, the steps of geo-spatial operation need to be clarified to LLM with texts. Corresponding codes in Python would help end-user to achieve their geospatial operations. One limitation of this work is that users are required to upload or download a prepared dataset, which restricts them from leveraging the vast amount of existing open geodata available on the Internet.

Therefore, in this work, we would like to present GeoQAMap, an evolving system designed to answer geospatial questions using maps. It is a further development of GPT-like AI system. We demonstrated a preliminary example integrating the state-of-the-art LLM and the public knowledge base Wikidata. The system follows a process where questions are first translated into SPARQL queries, which are then queried in a Wikidata endpoint. The output JSON is then utilized in conjunction with the Python library to create interactive maps that provide visual answers to the geographic questions.

¹ Stable Diffusion Online. Source: <https://stablediffusionweb.com/>



■ **Figure 2** For question “all universities of China and Germany”, the interface contains: prompt input box (upper left), the extracted name entities and generated SPARQL query (bottom left), additional context information (upper right), and the output map (bottom right).

2 Methodology

The entire workflow of the proposed GeoQAMap system is illustrated in Figure 1. There are in general three steps: (1) prompt optimization, (2) prompt interpretation and knowledge base query, and (3) map visualization. Figure 2 shows how this system interact with users.

2.1 Prompt optimization

Formulating the prompt sentence is essential as it directly influences the response generated by the LLM. Additionally, it is crucial to specify the desired output format. The majority of existing LLMs are not readily openly accessible to developers, limiting their ability to retrain or fine-tune the provided models together with the computational resource constraints. Therefore, in this work, frequently happened issues are recorded, and we improve the system performance by giving additional constraints in the prompt sentence.

Considering a user-submitted geospatial question in natural language, we have established several constraints for the output:

- (1) Specifically, we require the output to be a pure SPARQL sentence, devoid of any headers or explanatory text that could potentially create issues when interacting with the SPARQL endpoint.
- (2) According to our observations, the LLM system would often make mistakes on finding correct Wikidata ID for the corresponding name entities. Therefore, we ask the LLM first to extract name entities from users’ prompts, and then look up the corresponding Wikidata ID using the Wikidata API via HTTP requests.
- (3) Additionally, users are provided with an extra textbox to expand the prompt whenever they encounter incomplete results, allowing them to provide further context to refine their query. As in Figure 2, the extra context input make the LLM to consider the sub-classes (P279) of university in addition to instances (P31) when generating the query.

2.2 Prompt interpretation and query knowledge base

The GeoQAMap utilizes *GPT-3.5* as its underlying LLM, providing access to a range of natural language processing capabilities, including Name Entity Recognition (NER). To

28:4 GeoQAMap - Geographic Question Answering with Maps

interact with *GPT-3.5*, OpenAI API² was used. The generated SPARQL query would then be directly given to Wikidata Query Service, where the query is sent to the endpoint server³ via Python package *sparqlwrapper*.

Of course, there may be instances where the SPARQL query is not executable on the Wikidata Query Service. In such cases, users may need to manually intervene, for example, by utilizing the Wikidata Query Service web interface to verify the validity of the query sentence. This could involve checking for potential issues such as mismatched entity IDs, mismatched SPARQL syntax or other related issues. Even though this process may occasionally require user interaction, it still significantly reduces the effort compared to constructing complex SPARQL queries from scratch every time.

2.3 Visualization

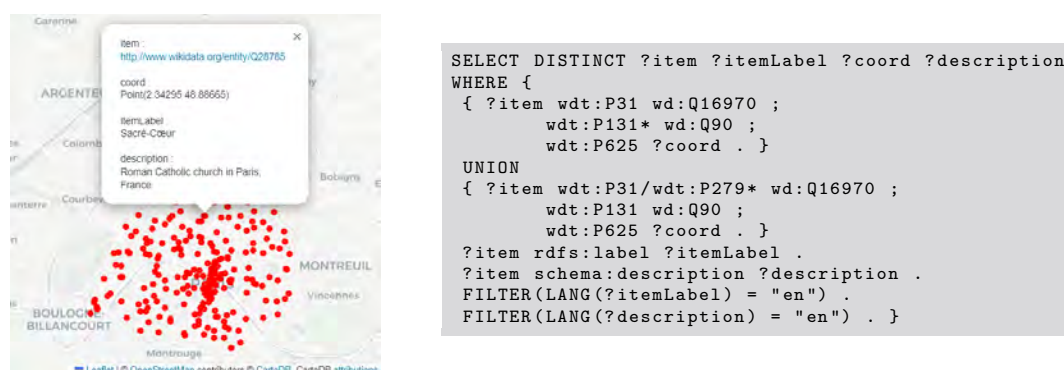
The Wikidata endpoint responses the SPARQL query with data in JSON format. The output text strings are then parsed and visualized with Python package *folium*. Within the standard Jupyter Notebook or Google Colab implemented with *gradio* interface, users can easily interact with this map and explore more details of the results.

3 Results and discussion

In this section, we present three case studies, which demonstrate the questions that GeoQAMap system can already answer with maps.

3.1 Questions for geo-entities of affiliation relationship

The most common type of questions is to search for specific geo-entities that are located within a certain administrative region. This type of query helps in finding relevant information about the relationship between geographical entities and the administrative regions they are associated with. With the first example, we present GeoQAMap's answer to the question "churches in Paris and all districts of Paris". The name entities of this question were first extracted with *church*, *Paris*, and *district*, where the corresponding Wikidata IDs are also given to the prompt sentence. The LLM-generated SPARQL query is in Figure 3.



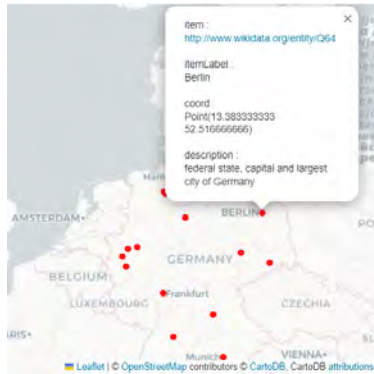
■ **Figure 3** Answering "churches in Paris and all districts of Paris".

² GPT - OpenAI API. Source: <https://platform.openai.com/docs/guides/gpt>

³ SPARQL endpoint server of Wikidata. Source: <https://query.wikidata.org/sparql>

3.2 Questions for geo-entities of attribute conditional filtering

A second frequent type of question is to select geo-entities with respect to certain criteria. With the second example, we would like to present GeoQAMap’s answer to the question “cities of Germany with a population more than 500,000”. Since many cities are only associated with the subclasses of “city (Q515)”, such as “big city (Q1549591)”, “Hanseatic city (Q707813)”. Therefore, the user would need to declare that “the subclasses of city should also be considered”. With the SPARQL generated as following, the answers as map in Figure 4 can be generated.



```

SELECT DISTINCT ?item ?itemLabel ?coord ?description
WHERE {
  ?item wdt:P31/wdt:P279* wd:Q515 .
  ?item wdt:P625 ?coord .
  ?item wdt:P17 wd:Q183 .
  ?item wdt:P1082 ?population .
  ?item rdfs:label ?itemLabel .
  ?item schema:description ?description .
  FILTER(LANG(?itemLabel) = "en" &&
    LANG(?description) = "en")
  FILTER(?population > 500000)
} ORDER BY ?population
  
```

■ **Figure 4** Answering “cities of Germany with a population more than 500,000”.

3.3 Questions for geo-entities that need further calculation

Moreover, some questions may need further calculation since the answers are not directly given in the Wikidata knowledge base. For example, a user queries “universities of the United Kingdom established more than 100 years”. Only the established time was recorded for universities in the Wikidata under the field of “inception (P571)”. However, this does not directly answer the user’s question. The LLM-generated SPARQL as in Figure 5 can perform the process of calculation properly. However, similar to the example in Figure 2, it would need to consider the subclasses of university. Therefore, in certain cases, a user may need to intervene and identify the errors to help the LLM to generate correct queries.



```

SELECT DISTINCT ?item ?itemLabel ?coord ?description
WHERE {
  ?item wdt:P31/wdt:P279* wd:Q3918;
  wdt:P625 ?coord;
  wdt:P17 wd:Q145.
  ?item wdt:P571 ?inception.
  FILTER(YEAR(NOW()) - YEAR(?inception) > 100)
  OPTIONAL {
    ?item schema:description ?description.
    FILTER(LANG(?description) = "en") }
  SERVICE wikibase:label { bd:serviceParam
    wikibase:language "[AUTO_LANGUAGE],en". }
} ORDER BY ?itemLabel
  
```

■ **Figure 5** Answering “universities of the United Kingdom established more than 100 years”.

3.4 Discussion

The system demonstrated in this work can already answer many geographic questions, especially questions such as the locations of geo-entities. Still, complicated questions that require geospatial operations, such as applying a buffer, are not yet achieved.

However, despite the advancements in automatically generating SPARQL queries, there are still several failure cases that often require manual intervention for correction. These failures can be broadly categorized into the following three types, as far as we observed:

- (1) Mismatch of named entities and relationships with incorrect Wikidata IDs: Although most cases can be resolved through extracting the named entities and look-up their code in the Wikidata server. It is reasonable to expect that the vast number of concepts and relationships in Wikidata may not be fully covered and learned by the language model.
- (2) Syntax errors: The LLM system may occasionally produce syntax errors, such as generating invalid syntax resulting in query inconsistencies. To address these issues, regular expression rules can be established to identify and rectify such syntax errors.
- (3) Inconsistencies in Wikidata knowledge base: Within Wikidata, geo-entities can be linked to different geographical entities, including nation names, city names, or city district names. While the general affiliation may be clear, there are instances where the associations are not accurately recorded in the Wikidata knowledge base. This discrepancy can lead to incorrect or incomplete results when querying geospatial information.

To handle these failure cases, manual intervention becomes necessary to identify errors in the generated SPARQL queries and communicate with the LLM to make it remember. Despite these challenges, the automated generation of SPARQL queries by LLMs still greatly reduces the overall effort required to construct complex queries from scratch. It serves as a valuable starting point, with manual correction acting as a backup step to refine and ensure the accuracy of the queries.

4 Conclusions and outlook

In this work, we presented our early implementation of GeoQAMap, a system that has been built on the current state-of-the-art LLM and open knowledge base to answer geospatial questions using maps. Many geospatial questions can be answered with an interactive map visualization and it allows users to explore details of individual geo-entities.

There are several aspects that require further exploration. Firstly, since there are still many cases that the LLM would generate incorrect SPARQL queries, it is important to comprehensively evaluate the performance of the current LLM models for this certain task and design proper strategies to ensure the correctness of the generated queries. Secondly, the implementation would benefit from the inclusion of a filter mechanism that determines which questions specifically require answers using maps and which associated geo-entities are in need of visualization for the user. Lastly, at present, the system's capabilities are limited to query-based questions, and the depth and breadth of its ability to answer complicated questions would require significant enhancements.

Furthermore, as illustrated in Figure 1 and highlighted in orange, we aim to leverage the capabilities of Virtual Knowledge Graph (VKG) technology to include more geospatial data into the process, e.g., OpenStreetMap and CityGML, in order to achieve geo-analytical question answering [4]. By combining Ontop⁴ and GeoSPARQL, Ding et al. (2021) [1]

⁴ Ontop - A Virtual Knowledge Graph System. Source: <https://ontop-vkg.org/>

demonstrated the ability to answer questions involving geospatial operations, such as buffering. The LLM can therefore act as a crucial entry point, allowing users to pose complex geospatial questions using natural language.

References

- 1 Linfang Ding, Guohui Xiao, Albulen Pano, Claus Stadler, and Diego Calvanese. Towards the next generation of the linkedgeodata project using virtual knowledge graphs. *Journal of Web Semantics*, 71:100662, 2021.
- 2 Yuhao Kang, Qianheng Zhang, and Robert Roth. The ethics of ai-generated maps: A study of dalle 2 and implications for cartography. *arXiv preprint arXiv:2304.10743*, 2023.
- 3 Zhenlong Li and Huan Ning. Autonomous gis: the next-generation ai-powered gis. *arXiv preprint arXiv:2305.06453*, 2023.
- 4 Simon Scheider, Enkhbold Nyamsuren, Han Kruiger, and Haiqi Xu. Geo-analytical question-answering with gis. *International Journal of Digital Earth*, 14(1):1–14, 2021.

Understanding the Complex Behaviours of Electric Vehicle Drivers with Agent-Based Models in Glasgow

Zixin Feng ✉

Urban Big Data Centre, School of Social and Political Sciences, University of Glasgow, UK

Qunshan Zhao ✉

Urban Big Data Centre, School of Social and Political Sciences, University of Glasgow, UK

Alison Heppenstall ✉

Urban Big Data Centre, School of Social and Political Sciences, University of Glasgow, UK

Abstract

With the new policy aimed at advancing the phase-out date for the sale of new petrol and diesel cars and vans to 2030, the electric vehicle (EV) market share is expected to rise significantly in the coming years. This necessitates a deeper understanding of the driving and charging behaviours of EV drivers to accurately estimate future charging demand distribution and benefit for future infrastructure development. Traditional data-based approaches are limited in illustrating the granular spatiotemporal dynamics of individuals. Recent studies that use conventional vehicle trajectory data also have the sampling bias problem, despite their analyses being conducted at a finer resolution. Moreover, studies that use simulation approaches are often either based on limited behaviour rules for EV drivers or implemented in an artificial grid environment, showing limitations in reflecting real-world situations. To address the challenges, this work introduces an agent-based model (ABM) with complex behaviour rules for EV drivers, taking into account the drivers' sensitivities to financial and time costs, as well as route deviation. By integrating the simulation model with the origin and destination information of drivers, this work can contribute to a better understanding of the behaviour patterns of EV drivers.

2012 ACM Subject Classification Computing methodologies → Modeling and simulation

Keywords and phrases Electric vehicles, agent-based modelling, charging demand, route choices

Digital Object Identifier 10.4230/LIPIcs.GIScience.2023.29

Category Short Paper

Funding The work was made possible by the ESRC's on-going support for the Urban Big Data Centre (UBDC) [ES/L011921/1 and ES/S007105/1].

1 Introduction

The transition from petrol/diesel-based vehicles to alternative fuel vehicles can play an important role in reducing global greenhouse gas emissions and air pollution. The UK government has announced to bring forward the phase-out date for the sale of new petrol and diesel cars and vans to 2030, and to require all new cars and vans to be completely zero-emission at the tailpipe from 2035. The new policy targets necessitate a higher electric vehicle (EV) penetration rate and create opportunities in the market for electric vehicles. Therefore, it is essential to illustrate the behaviour patterns of EV drivers and provide a deeper understanding of the spatial distribution of their charging demand.

Data-driven approaches have been used to explore EV driver behaviour and charging demand in previous research, mostly relying on statistical methods to understand charging behaviours [7]. However, the use of socio-demographic statistics [5] and travel survey data [2],



© Zixin Feng, Qunshan Zhao, and Alison Heppenstall;
licensed under Creative Commons License CC-BY 4.0

12th International Conference on Geographic Information Science (GIScience 2023).

Editors: Roger Beecham, Jed A. Long, Dianna Smith, Qunshan Zhao, and Sarah Wise; Article No. 29; pp. 29:1–29:6

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

■ **Table 1** The attributes of hypothetical charging stations.

Station name	Charger name	Charge speed(kW)	Charge price (£/kWh)
Station 1	1A	6	0.3
Station 1	1B	2	0.1
Station 2	2	12	0.6

despite their rich attributes, can be limited in demonstrating the granular spatiotemporal dynamics of individuals and their decision-making processes. Alternatively, recent studies have used GPS data from conventional vehicles [9] or EV fleets like taxis [12] to understand EV behaviours. Although these datasets can demonstrate spatiotemporal trajectories of drivers in a finer resolution, their ability to represent all types of EVs – including private EVs, ride-hailing EVs, and other commercial EV fleets – can be questionable. As a result, these datasets could introduce sampling bias and lead to systematic errors in the conclusions.

More importantly, the utilisation of the datasets above cannot fully capture the various features of EV drivers' behaviours, including their sensitivities to charging costs and psychological preferences. Sensitivities to charging costs can affect a driver's behaviours in various ways, such as travel distance [13], the payment for charging [6], and the time spent waiting and charging [3, 14]. Meanwhile, the psychological factor refers to a driver's comfort level with a low State of Charge (SOC) [15]. Given the limited availability of EV trajectory data, integrating these complex EV driver behaviour rules with the origin and destination (OD) information of drivers through simulation methods can provide opportunities to explore detailed EV trajectories and a granular charging demand distribution.

Agent-based modelling (ABM) offers a simulation method to plan, design, and experiment with micro-agents in an artificial computational environment [11]. Compared to statistical methods, ABMs can represent a richer and more detailed set of individual agents [4] and enable interactions both between and within agent types [8]. Previous studies have applied ABM to simulate the behaviours of EV drivers. However, research gaps exist because they either captured limited behavioural rules of EV drivers [16], or were implemented in hypothetical grid environments rather than real-world road networks [1]. Consequently, this work aims to provide a deeper understanding of the driving and charging behaviours of EV drivers by creating an ABM with comprehensive behaviour rules and implementing the model in a real-world road network in Glasgow.

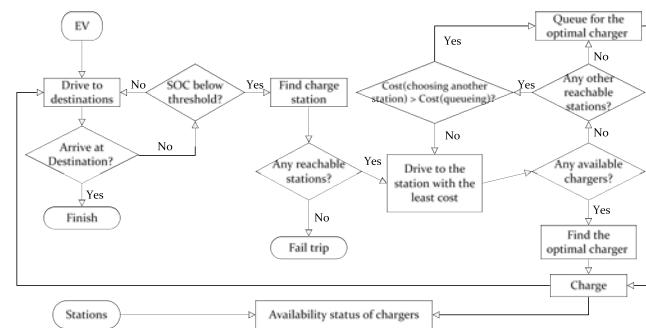
2 Data and methods

2.1 Data and study area

The geographical scale covers the area between Glasgow city centre and the West End in Glasgow. The study area is shown in Figure 2. As described in Table 1, two hypothetical charging stations are situated within the area. Station 1 is equipped with two chargers, while Station 2 has one charger. Each charging station operates at a hypothetical charging speed and charging price. The origins and destinations of EV drivers are stochastically selected from the 869 nodes on the road network using a simple random sampling method. Twenty-two cars are simulated in each iteration.

2.2 Model development

In the model, the SOC consumption rate is assumed to be constant, as operating conditions and environmental influences that can potentially affect the energy usage of vehicles are not considered. Furthermore, a driver is sensitive to a particular SOC threshold, meaning that when the SOC falls below this threshold, the driver recognizes that the battery has depleted to a level where recharging is necessary. The full battery capacity of the vehicles is assumed to be 100kWh [10]. The starting SOC value of each driver is set at 100%. The SOC threshold is randomly selected to be either 40% or 50%. This threshold ensures a considerable amount of charge remains in the battery, thus reducing the likelihood that the driver will be stranded without power for the simulation purposes. An ABM model is developed using the Mesa package in Python to integrate the complex behaviours of EV drivers. The model includes two types of agents: EV drivers and charging stations. The behavioural rules of the model are demonstrated in Figure 1.



■ **Figure 1** Behavioural rules of EV drivers and charging stations.

Drivers start their journeys from the origins and drive to the destinations following the Dijkstra's shortest path on the road network. A driver searches for an optimal charging station with least cost when the SOC falls below the SOC threshold. The cost of station p is denoted by equation 1. α represents a driver's sensitivity to charge payment. β , ϵ and δ represent the cost in GBP that a driver attributes to each additional unit of travel distance to find a charging station (GBP/km), each extra unit of time to spend while charging (GBP/minute), and each decrease in the degree of availability at a station (GBP/%), respectively. Since the drivers cannot predict the queueing time at the station before their arrival, we use *availability*, a value ranging from 0 to 1, to indicate the percentage of available chargers at the station. This provides an estimate of the likelihood of not encountering a queue upon arrival at the station.

$$Cost_{station_p} = \alpha \times payment_p + \beta \times distance_p + \epsilon \times charge_time_p + \delta \times availability_p \quad (1)$$

If a driver cannot find a station within reach based on the current SOC, the status of the EV driver changes to 'fail trip', and the driver will not be able to continue the journey. Otherwise, the driver will follow the shortest path to the optimal charging station. Upon arrival at the station, drivers are updated with the most current availability status of the chargers. If at least one charger is available, the driver will select the optimally available charger with the least cost. If all the chargers are occupied and another station is reachable given the current SOC status, the EV driver will compare the cost of queuing to the cost of finding another station, choosing the option with the lower cost. Otherwise, the driver will queue for the optimal charger before beginning to charge. The cost of charger q at the

29:4 Simulating Electric Vehicle Driver Behaviours with ABM

selected station is denoted by equation 2. It comprises the charge payment and the cost in GBP that a driver attributes to the time spent queueing and charging. It should be noted that the payment and charge time in equation (2) can be different from equation (1) due to the change of charging circumstances in the charging station when the drivers arrive.

$$Cost_{charger_q} = \alpha \times payment_q + \epsilon \times charge_time_q + \epsilon \times queue_time_q \quad (2)$$

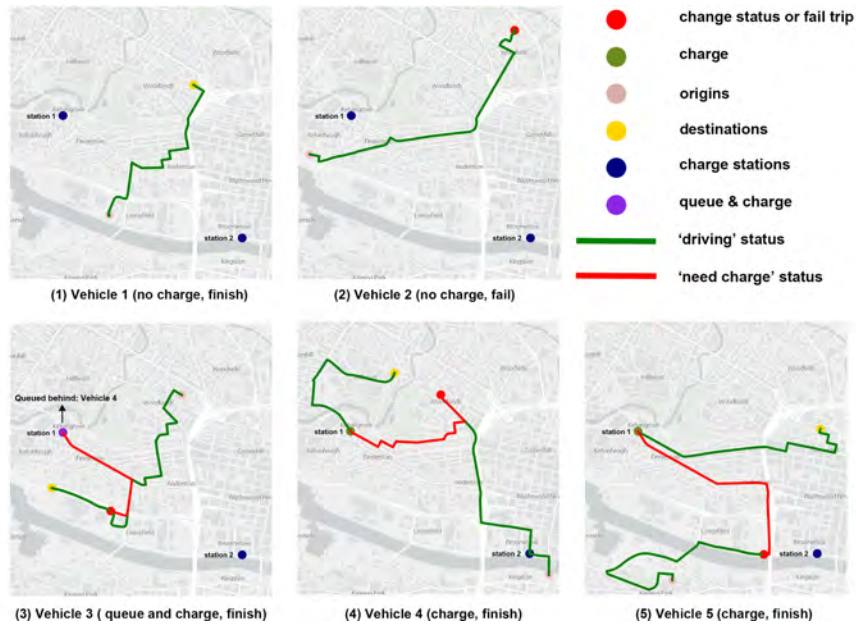
If the SOC drops below the threshold again, the driver will recharge. Upon reaching the destination, the driver evaluates the total cost incurred during the trip ($Total_cost$), which is calculated in equation 4. It is composed of the deviation and the cost spent at all used chargers. $Deviation$ is calculated in equation 3. It refers to the difference in distance between the planned route (the shortest path between the origin and destination) and the actual route taken by the driver. γ represents the cost in GBP that a driver attributes to each unit of deviation (GBP/km). i is the i -th decision made by the driver while j is the j -th charger used by the driver. m is the total number of decisions made by a driver and n is the total number of chargers used by a driver.

$$Deviation = \sum_{i=1}^{m-1} Distance_{i,i+1} - Distance_{plan} \quad (3)$$

$$Total_cost = \gamma \times Deviation + \sum_{j=1}^n Cost_{charger_j} \quad (4)$$

3 Results and discussion

3.1 Simulation Results



■ Figure 2 EV routes.

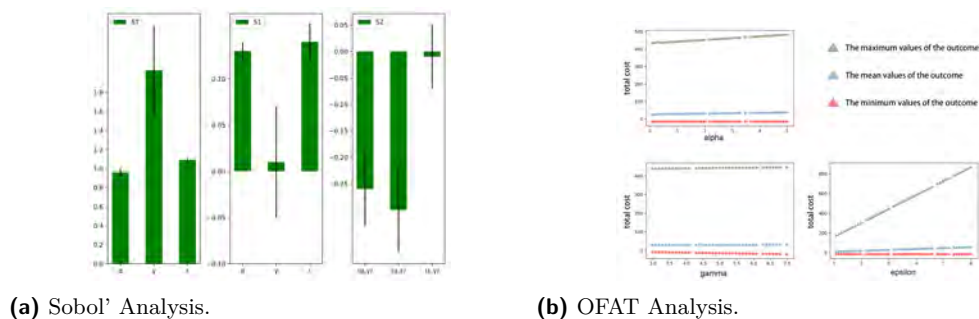
■ **Table 2** Simulation results.

Vehicle ID	Final status	Deviation (km)	Cost (£)	Charge time (min)	Charger	Queue time (min)
1	finish	0	0	-	-	0
2	fail trip	0	0	-	-	0
3	finish	1.88	9.95	[51.43,150.94]	1B	41.48
4	finish	0.56	8.57	[8.57, 51.43]	1B	0
5	finish	0.24	28.7	[9.58, 25.54]	1A	0

Five samples of the simulation results are presented in Table 2. The visualised routes are shown in Figure 2. EV-2 fails to complete its journey because it cannot reach any of the charging stations with the remaining SOC. EV-3, EV-4, and EV-5 charged en route. EV-3 queued behind EV-4 for charger “1B”, resulting in a wait time of 41.48 minutes.

3.2 Model calibration

As shown in Figure 3, the One Factor at a Time (OFAT) and Sobol’ method-based sensitivity analyses were performed to determine the robustness of the results and guide further improvements of the model. The confidence interval was calculated at 95%. The results suggest the presence of higher-order interactions in the function, as the total-order indices are larger than the first-order indices for all parameters. Furthermore, the OFAT sensitivity analysis is conducted to explore the sensitivity of total cost to charge payments (α), to deviation (γ), and to charge and queue time (ϵ). The variations of α , γ , and ϵ can affect the total cost if a driver charges during the trip, but have no effect on the total cost if a driver does not charge (Total cost = 0).



■ **Figure 3** Sensitivity analysis results.

4 Conclusion and future works

This work has developed an ABM that integrates the complex behaviours of EV drivers. The simulation results show that the drivers’ sensitivity to deviation is the strongest determinant of the total cost. Additionally, the results of the sensitivity analysis show that the cost function needs to be further modified to include nonlinear terms and interaction terms between the variables.

The model presented is based on multiple assumptions and is limited in reflecting real-world situations. In future work, we plan to replace the hypothetical stations with real-world public charging stations, and substitute the randomly generated OD information with real-world trajectory data of vehicles. This will enable us to explore how driving patterns might change when a driver adopts an EV. Based on the trajectory data, we also aim to integrate

the heterogeneous behavioural rules of different drivers into the model. Charging session data will also be used to validate the simulation results and ensure that the simulated driving and charging behaviours accurately represent real-world scenarios.


References

- 1 Kalpesh Chaudhari, Nandha Kumar Kandasamy, Ashok Krishnan, Abhisek Ukil, and Hoay Beng Gooi. Agent-based aggregated behavior modeling for electric vehicle charging load. *IEEE Transactions on Industrial Informatics*, 15(2):856–868, 2018.
- 2 Constance Crozier, Thomas Morstyn, and Malcolm McCulloch. Capturing diversity in electric vehicle charging behaviour for network capacity estimation. *Transportation Research Part D: Transport and Environment*, 93:102762, 2021.
- 3 Sreten Davidov. Optimal charging infrastructure planning based on a charging convenience buffer. *Energy*, 192:116655, 2020.
- 4 Gary A Davis and Paul Morris. Statistical versus simulation models in safety: steps toward a synthesis using median-crossing crashes. *Transportation research record*, 2102(1):93–100, 2009.
- 5 Guanpeng Dong, Jing Ma, Ran Wei, and Jonathan Haycox. Electric vehicle charging point placement optimisation by exploiting spatial statistics and maximal coverage location models. *Transportation Research Part D: Transport and Environment*, 67:77–88, 2019.
- 6 Yanbo Ge, Don MacKenzie, and David R Keith. Gas anxiety and the charging choices of plug-in hybrid electric vehicle drivers. *Transportation Research Part D: Transport and Environment*, 64:111–121, 2018.
- 7 Jurjen R Helmus, Michael H Lees, and Robert van den Hoed. A data driven typology of electric vehicle user types and charging sessions. *Transportation Research Part C: Emerging Technologies*, 115:102637, 2020.
- 8 Alison Heppenstall, Andrew Crooks, Nick Malleson, Ed Manley, Jiaqi Ge, and Michael Batty. Future developments in geographical agent-based models: Challenges and opportunities. *Geographical Analysis*, 53(1):76–91, 2021.
- 9 Eleftheria Kontou, Changzheng Liu, Fei Xie, Xing Wu, and Zhenhong Lin. Understanding the linkage between electric vehicle charging network coverage and charging opportunity using gps travel data. *Transportation Research Part C: Emerging Technologies*, 98:1–13, 2019.
- 10 Lizi Luo, Wei Gu, Suyang Zhou, He Huang, Song Gao, Jun Han, Zhi Wu, and Xiaobo Dou. Optimal planning of electric vehicle charging stations comprising multi-types of charging facilities. *Applied energy*, 226:1087–1099, 2018.
- 11 Sedar Olmez, Jason Thompson, Ellie Marfleet, Keiran Suchak, Alison Heppenstall, Ed Manley, Annabel Whipp, and Rajith Vidanaarachchi. An agent-based model of heterogeneous driver behaviour and its impact on energy consumption and costs in urban space. *Energies*, 15(11):4031, 2022.
- 12 Wei Tu, Qingquan Li, Zhixiang Fang, Shih-lung Shaw, Baoding Zhou, and Xiaomeng Chang. Optimizing the locations of electric taxi charging stations: A spatial-temporal demand coverage approach. *Transportation Research Part C: Emerging Technologies*, 65:172–189, 2016.
- 13 Rick Wolbertus, Robert van den Hoed, Maarten Kroesen, and Caspar Chorus. Charging infrastructure roll-out strategies for large scale introduction of electric vehicles in urban areas: An agent-based simulation study. *Transportation Research Part A: Policy and Practice*, 148:262–285, 2021.
- 14 Yang Yang, Enjian Yao, Zhiqiang Yang, and Rui Zhang. Modeling the charging and route choice behavior of bev drivers. *Transportation Research Part C: Emerging Technologies*, 65:190–204, 2016.
- 15 Tao Yi, Chao Zhang, Tongyao Lin, and Jinpeng Liu. Research on the spatial-temporal distribution of electric vehicle charging load demand: A case study in china. *Journal of Cleaner Production*, 242:118457, 2020.
- 16 Zhiyan Yi, Bingkun Chen, Xiaoyue Cathy Liu, Ran Wei, Jianli Chen, and Zhuo Chen. An agent-based modeling approach for public charging demand estimation and charging station location optimization at urban scale. *Computers, Environment and Urban Systems*, 101:101949, 2023.

Progress in Constructing an Open Map Generalization Data Set for Deep Learning

Cheng Fu ✉ 

Department of Geography, University of Zürich, Switzerland

Zhiyong Zhou ✉ 

Department of Geography, University of Zürich, Switzerland

Jan Winkler ✉ 

Department of Environmental Systems Science, Swiss Federal Institute of Technology, Zürich, Switzerland

Nicolas Beglinger ✉

swisstopo, Swiss Federal Office of Topography, Wabern, Switzerland

Robert Weibel ✉ 

Department of Geography, University of Zürich, Switzerland

Abstract

Recent pioneering works have shown the potential of a new deep-learning-backed paradigm for automated map generalization. However, this approach also puts a high demand on the availability of balanced and rich training sets. We present our design and progress of constructing an open training data set that can support relevant studies, collaborating with the Swiss Federal Office of Topography. The proposed data set will contain transitions of building and road generalization in Swiss maps at 1:25k, 1:50k, and 1:100k. By analyzing the generalization operators involved in these transitions, we also propose several challenges that can benefit from our proposed data set. Besides, we hope to also stimulate the production of further open data sets for deep-learning-backed map generalization.

2012 ACM Subject Classification Information systems → Geographic information systems; Information systems → Data mining; Information systems → Document structure

Keywords and phrases open data, deep learning, map generalization

Digital Object Identifier 10.4230/LIPICs.GIScience.2023.30

Category Short Paper

Funding This work has received funding from the Swiss National Science Foundation under project number 200021_204081, project DeepGeneralization.

Acknowledgements We would like to thank Roman Geisthoewel at swisstopo for his kind support and helpful discussion.

1 Introduction

Map generalization is a cartographic process for deriving a target map or database at a reduced scale from a source database by reducing the contents and complexity of the map while preserving necessary information of the map at the source scale [12]. Despite a long history of attempts to develop a fully automated pipeline with the assistance of machine learning [13, 9], map generalization still requires significant manual intervention by expert cartographers. The recent success of deep learning (DL, [8]) in computer vision has led researchers in cartography to adapt DL models toward an end-to-end map generalization workflow. Current studies focus on the generalization of buildings [16, 5] and roads [3, 4] in transitions between large scales, e.g. 1:5k to 1:20k, using raster- or vector-based data models.



© Cheng Fu, Zhiyong Zhou, Jan Winkler, Nicolas Beglinger, and Robert Weibel; licensed under Creative Commons License CC-BY 4.0

12th International Conference on Geographic Information Science (GIScience 2023).

Editors: Roger Beecham, Jed A. Long, Dianna Smith, Qunshan Zhao, and Sarah Wise; Article No. 30; pp. 30:1–30:6

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

The success of DL models exploits the increased complexity of neural network infrastructures to increase the learning capacity, but it also needs the support of big data. In the early stage of DL in computer vision applications, big open data sets such as ImageNet [7] contributed to the development of models by serving as baselines and allowing researchers to focus on improving the methods. Similarly, applying DL in map generalization also needs the support of big data. Unlike classical computer vision tasks (e.g., semantic segmentation and instance segmentation) that solely focus on individual objects, map generalization mainly targets the global organization of geographic objects, which is comprehensively influenced by their geometric and semantic characteristics. Besides, different scales involve different generalization criteria [10]. Therefore, training sets for DL applications in map generalization need extra effort, compared to computer vision.

To promote the progress of automated map generalization models in the era of deep learning, we set out to construct an open data set for map generalization as one of the major tasks in the DeepGeneralization project with support from swisstopo, the Swiss Federal Office of Topography. This report presents the design and the most recent progress in implementing the data set.

2 Design

2.1 Raw data and scope

The raw data sets contributed by swisstopo include KRM_25 (cartographic reference model at 1:25k; KRM: *Kartografisches Referenzmodell* in German) and DKM_25/50/100 (digital cartographic model at 1:25k/50k/100k, respectively; DKM: *Digitales Kartografisches Modell* in German). KRM is directly derived from swisstopo's Topographic Landscape Model (TLM) without much geometric adjustment. DKMs are the generalized geometries of the KRM that end up in the final map products. Besides essential information such as the geometry and other necessary attributes, each entity in each raw data set has a UUID to trace the possible transformation between maps. A join table is applicable to trace the changes between two consecutive scales, such as aggregation for the generalization between two maps.

While cartographers at swisstopo have a well-documented and well-established workflow for map generalization, the matching of UUIDs between maps of two consecutive scales is not guaranteed. A missing UUID on a smaller-scale map might be the result of deletion or aggregation. It is not always a reduction in UUIDs, as a generalized map may have new geometric entities that do not exist in the source map, due to cartographic reasons. For example, a road with two segments in the 1:25k map may have three segments in the 1:50k map. In addition, there is no information regarding the generalization operators applied. Therefore, matching is still needed to link the records in different maps, especially based on the spatial relationships among the geometries.

A balanced training set is critical for machine-learning models. In the context of map generalization, the balance can regard the instances of different map generalization operators, the spatial contexts/constraints between buildings and roads, land use contexts such as urban vs. rural, etc. Following the recent rise of explainable AI (XAI, [6]), researchers may also want to evaluate if one network structure works better for one map generalization task than another or how the network performs for a specific operator. Researchers thus may want to build up their own sampling strategies to balance the instances based on specific operators or geometric metrics. Thus, to better serve the community, only clarifying the corresponding relationship between a source entity and its target entity in the generalized map is not enough. We decided to include the map operator descriptions as part of the metadata for the map generalization cases.

Based on current research priorities, the planned data set will cover the transformations of buildings and roads between the three scales with vector-based outputs, from which raster-based representations can be easily derived.

2.2 A conceptual model for map generalization transformations

The transformations between the three scales we are using mainly include the generalization operators selection (elimination), simplification, aggregation, displacement, exaggeration, typification, and smoothing. We categorize the operators into atomic operators, including all aforementioned operators except typification, which is categorized as a complex operator that consists of atomic operators. Our classification is based on cartographic knowledge and the cardinality between the source and target geometries: An atomic operator involves 1 : 1 or $N : 1$ relationships, while complex operators such as typification usually involve $N : M$ relationships. The complex operators are hard to characterize formally, as they involve many scenarios for which even professional cartographers may not reach a consensus.

Based on the conceptual model, we thus formalize the operators in a generalization transformation between a source geometry and a target geometry as a series of selected atomic operators using a set $T \subset \{deletion, simplification, displacement, aggregation, exaggeration, smoothing\}$. A complete transformation thus consists of the source and target entities and the operator set for each pair of source and target entities.

2.3 An automated workflow

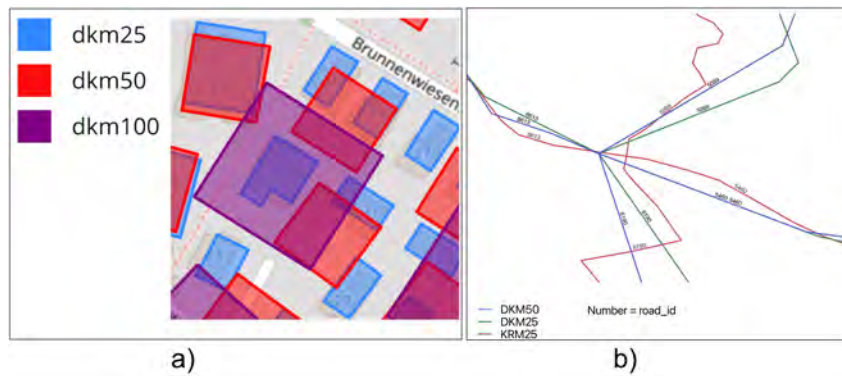
To derive corresponding source-target pairs and the applied generalization operators from the raw data sets, an automated workflow was designed, as manual matching is impossible due to the large number of samples.

The matching workflows for buildings and roads are performed separately, though both start with the UUID-join table. For building matching and operator detection, an additional spatial join was applied to the source and target buildings to find intersecting pairs. With the intersection table, *aggregation* was determined if a target building spatially overlaps with more than one source building. *Displacement* was detected if centroids of buildings with the same UUID exceed a buffer distance. *Simplification* was identified based on the change in terms of shape complexity [2]. Only entities not being part of an aggregation were fed into the *enlargement* detection module, as we regard the aggregated buildings as new, synthetic entities. *Deletion* was determined if a UUID was removed from the target map. All modules exported the decision as a binary result, which form a 5-dimension map-operator vector.

The workflow for determining generalization operators on roads is more challenging. Currently, we are still developing the matching module. The reason is that roads involve more complex geometric changes compared to the buildings, such as Figure 1. The spatial relationships between road segments cannot be simply inferred by intersection detection. The matching also relies on proper distances to describe the similarity between two lines. We chose the number of vertices, curviness, and sinuosity as the main metrics to characterize roads. However, how to determine the map operators based on the transformation of metrics between maps with different scales still need further conceptualization and development.

2.4 Database schema

The proposed data set will be delivered as two loosely connected databases: A Postgres database will store the geometry, UUID, and other attributes from the raw swisstopo data set. A MongoDB database will be used to store the transformation information of entity pairs,

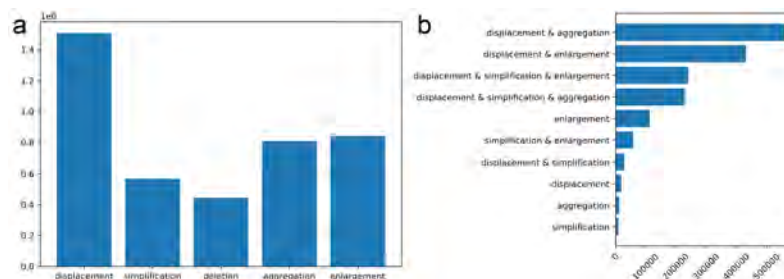


■ **Figure 1** a. An example of buildings; b. An example of roads at an intersection.

with individual collections for buildings and roads. We chose a NoSQL solution because transformation information differs case by case, while MongoDB has minimal data structure constraints. Each collection will contain the associated UUIDs with modeled operator types and metrics to characterize the transformation between the two entities. The collection will also have metrics extracted from the geometries, such as the number of vertices and shape complexity, which can benefit the data set users to design their own sampling strategies for compiling a customized training data set.

3 Constructing progress

Our workflow for buildings is well established and was applied to transitions between 1:25k and 1:50k maps in Switzerland, in which the source and the target map are both at medium scales, for preliminary testing. It can be observed that most of the building geometries are displaced after the generalization (Figure 2.a). Cases with only a single operator are rare. The instances of different combinations of the automated map operators are also highly imbalanced (Figure 2.b), suggesting that learning the implicit map generalization rules can be challenging.



■ **Figure 2** Map generalization operator cases of 2,078,548 building entities in 1:25k to 1:50k. a. By operator type; b. By operator combinations.

4 Research agenda

Challenge 1: Learning dominant but neglected operators

Using deep learning to explicitly learn individual generalization operators is mainly based on vector maps, which can reduce the manual intervention of expert cartographers (e.g., for setting thresholds). As illustrated in Figure 2.a, *displacement* is the predominant operator involved in map generalization within medium scales, followed by the *enlargement*, *aggregation*, *simplification*, and *deletion* operators. Unfortunately, it seems that the more dominant operators, including *displacement* and *enlargement*, are paid less attention to while some studies have attempted to learn *aggregation* [15], *simplification* [16], and *deletion* [14]. Therefore, research efforts should be particularly directed towards learning *displacement* and *enlargement* by formulating them as learnable tasks and introducing feasible models.

Challenge 2: Developing end-to-end generalization models

While the learning of individual generalization operators benefits the explicit modeling of cartographic knowledge and achieves the generalization of a part of map objects, it is still necessary to chain these intermediate outputs for more map objects to produce the final generalized map. Therefore, a second stream of raster-based deep learning models has great potential to enable end-to-end map generalization [4, 5]. The existing studies mainly work on the *aggregation*, *simplification*, and *deletion* operators and their combinations [5]. However, Figure 2.b shows that map generalization for medium scales also involves a large portion of combinations of *displacement* and other operators, as well as further, different combinations. Therefore, the raster-based deep learning models should be further developed using a more comprehensive data set that contains the dominant operator combinations (e.g., our demonstrated swisstopo data set) to improve their capacity for end-to-end solutions.

Challenge 3: Understanding learned cartographic knowledge

While Challenge 1 is result-oriented, it is also important to understand what specific cartographic knowledge a DL network learns. From the pragmatic perspective, this can contribute to better fine-tuning strategies for learning; from the theoretical perspective, it helps to gain scientific knowledge on the capacity and limitations of DL network architectures. XAI methods such as Grad-Cam [11] for raster-based data and those in GraphXAI [1] for vector-based graphs can help gaining interpretation of the knowledge a DL network learns. The result can also guide the optimization or chaining of modules in an end-to-end generalization workflow in Challenge 2, if the final generalization turns out to be a cascading multi-module workflow.

5 Future work

In future work, we would like to take a closer look at the validation of the generalization operator modeling and its effectiveness on the 1:50k to 1:100k generalization of buildings, with the help of swisstopo cartographers. The generalization operator modeling for roads will continue. Once the data set is published, a crowd-sourcing-like approach may also be applied for collecting corrections of specific transformations.

References

- 1 Chirag Agarwal, Owen Queen, Himabindu Lakkaraju, and Marinka Zitnik. Evaluating explainability for graph neural networks. *Scientific Data*, 10(144), 2023.
- 2 Thomas Brinkhoff, Hans-Peter Kriegel, Ralf Schneider, and Alexander Braun. Measuring the complexity of polygonal objects. In *ACM-GIS*, volume 109, 1995.
- 3 Azelle Courtial, Guillaume Touya, and Xiang Zhang. Constraint-based evaluation of map images generalized by deep learning. *Journal of Geovisualization and Spatial Analysis*, 6(1):13, 2022.
- 4 Azelle Courtial, Guillaume Touya, and Xiang Zhang. Deriving map images of generalised mountain roads with generative adversarial networks. *International Journal of Geographical Information Science*, 37(3):499–528, 2023.
- 5 Yu Feng, Frank Thiemann, and Monika Sester. Learning cartographic building generalization with deep convolutional neural networks. *ISPRS International Journal of Geo-Information*, 8(6):258, 2019.
- 6 David Gunning, Mark Stefk, Jaesik Choi, Timothy Miller, Simone Stumpf, and Guang-Zhong Yang. Xai—explainable artificial intelligence. *Science robotics*, 4(37):eaay7120, 2019.
- 7 Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE, June 2009.
- 8 Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.
- 9 Corinne Plazanet, Nara Martini Bigolin, and Anne Ruas. Experiments with learning techniques for spatial model enrichment and line generalization. *GeoInformatica*, 2:315–333, 1998.
- 10 L Tiina Sarjakoski. Conceptual models of generalisation and multiple representation. *Generalisation of Geographic Information*, pages 11–35, 2007.
- 11 Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.
- 12 Robert Weibel and Geoffrey Dutton. Generalising spatial data and dealing with multiple representations. In P.A. Longley, M.F. Goodchild, D.J. Maguire, and D. Rhind, editors, *Geographical Information Systems: Principles and Applications*, volume 1, pages 125–155. Longman Scientific & Technical, London, 1999.
- 13 Robert Weibel, Stefan Keller, and Tumasch Reichenbacher. Overcoming the knowledge acquisition bottleneck in map generalization: The role of interactive systems and computational intelligence. In Andrew U. Frank and Werner Kuhn, editors, *Spatial Information Theory A Theoretical Basis for GIS*, volume 988, pages 139–156, Berlin, Heidelberg, 1995. Springer Berlin Heidelberg.
- 14 Tianyuan Xiao, Tinghua Ai, Huafei Yu, Min Yang, and Pengcheng Liu. A point selection method in map generalization using graph convolutional network model. *Cartography and Geographic Information Science*, pages 1–21, 2023.
- 15 Xiongfeng Yan, Tinghua Ai, Min Yang, and Hongmei Yin. A graph convolutional neural network for classification of building patterns using spatial vector data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 150:259–273, 2019.
- 16 Zhiyong Zhou, Cheng Fu, and Robert Weibel. Move and remove: Multi-task learning for building simplification in vector maps with a graph convolutional neural network. *ISPRS Journal of Photogrammetry and Remote Sensing*, 202:205–218, 2023.

Project-Based Urban Dynamics: A Novel Method for Assessing Urban Sprawl

Nir Fulman ✉ 🏠

Department of Geography and Human Environment, Porter School of Environmental Studies, Tel Aviv University, Israel

GIScience Research Group, Heidelberg University, Germany

Yulia Grinblat

Heidelberg Institute for Geoinformation Technology (HeiGIT) gGmbH at Heidelberg University, Germany

Itzhak Benenson

Department of Geography and Human Environment, Porter School of Environmental Studies, Tel Aviv University, Israel

Abstract

We present a new approach to categorizing different types of urban development, namely infilling, fringe, and leapfrogging, based on construction projects as the fundamental unit of analysis. We focus on the role of the leapfrogging projects as seeds for new developments, leading to urban sprawl extending beyond statutory plans. To examine this phenomenon, we analyze the 50-year growth of three major Israeli cities: Netanya, Haifa, and Safed and the 5-year dynamics of 66 cities in Israel that account for 68% of the country's population. Our investigation utilizes extensive databases of Israeli development plans, along with high-resolution aerial photographs covering the investigated areas and time periods. These datasets were supplemented by detailed Israeli databases encompassing roads, buildings, and other infrastructure elements, compiled by the Israeli Mapping Centre for the year 2018. Our analysis reveals that although most construction projects in Israel adhere to land-use plans, urban sprawl in Israel remains highly unpredictable. Leapfrogging is specific in terms of both place and time, attracts additional development nearby, and forces the divergence from development plans. We conclude that urban modelers' view of urban dynamics being driven by common and systematic forces, is unrealistic. Instead, every city has its specific and self-enforcing development drivers that define its land-use dynamics. This explains the limited success of the Cellular Automata (CA) models in explaining and predicting urban dynamics.

2012 ACM Subject Classification Computing methodologies

Keywords and phrases Urban sprawl, Leapfrogging, GIS analysis, Complex system

Digital Object Identifier 10.4230/LIPICs.GIScience.2023.31

Category Short Paper

1 Introduction

Urban development is complex and only partially predictable, as illustrated by the limited ability of Cellular Automata (CA) to predict Land-Use/Land-Cover (LULC) dynamics ([3]). This is particularly true for leapfrog development beyond the current city boundary ([2]). Leapfrogging attracts additional development, and this positive feedback mechanism may override statutory plans ([1]) and significantly modify the city's spatial dynamics ([4]), increasing their unpredictability. To mitigate deviations from the development plans, it is crucially important to estimate the role of leapfrogging in urban dynamics. Our paper proposes a new method for identifying leapfrogging and assessing its effects by studying a large database of Israeli development plans versus real development.



© Nir Fulman, Yulia Grinblat, and Itzhak Benenson;
licensed under Creative Commons License CC-BY 4.0

12th International Conference on Geographic Information Science (GIScience 2023).

Editors: Roger Beecham, Jed A. Long, Dianna Smith, Qunshan Zhao, and Sarah Wise; Article No. 31; pp. 31:1–31:6

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

We depart from the conventional raster-based analysis of satellite images by analyzing urban sprawl based on the fundamental unit of urban development – the development project. Our view of urban dynamics centers on three types of urban development - infilling, fringe, and leapfrogging. We quantify the extent and attractiveness of leapfrog projects for further construction nearby that contributes to non-planned sprawl. Our study exploits unique county-wide Israeli data on land-use dynamics: aerial photographs, development plans starting from the 1960s, and comprehensive databases of building footprints, land use, and road data, all provided by the Israeli Mapping Center (IMC). The research focuses on a 53-year sprawl of three Israeli cities – Netanya, Haifa, and Safed – from 1964 to 2018, and the sprawl in 66 Israeli cities with a population exceeding 15,000, between 2013 and 2018.

2 The data

To assess the effects of leapfrogging, we investigate two datasets. The first represents long-term dynamics in three cities that differ in their properties: Haifa, a metropolitan city with a population of 283,000 in 2018; Netanya, a mid-sized city near Tel Aviv (217,000); and Safed, a small city located far from metropolitan areas (36,000). In each city, we study the LULC dynamics of 6-km width transects that start in the city’s CBD and extend beyond city boundaries to open spaces. The second dataset represents LULC dynamics in 66 cities housing 68% of Israel’s population, between 2013 and 2018.

Aerial photos covering the transects at a spatial resolution of 25 cm, were obtained from the IMC for the years 1964, 1972, 1983, 1993, 2000, and 2008. Based on each photo, polygons of building constructions and roads were manually digitized. Buildings and road layers for 2013 and 2018 were obtained from the IMC database for these years. To estimate the year of building construction we compared the IMC layer of buildings in 2018 to the corresponding aerial photography building layer, and assigned the year in which a building first appeared in the aerial photo as its construction year. The information on the building’s use was also acquired from the IMC layer for 2018 and aggregated into residential, industrial, public, and others. Additionally, we used the IMC 2018 road layer to estimate the year of road construction. All these data were matched to layers of construction plans. Similarly, we matched the IMC layers and development plans for the years 2013 and 2018 for the 66 cities.

Our assessment of the leapfrogging is based on the recognition of the urban fringe – the border area between the built-up and non-built-up parts of the city, and development projects – the basic units of urban development that consist of one or several buildings.

3 Identifying the urban fringe and development project

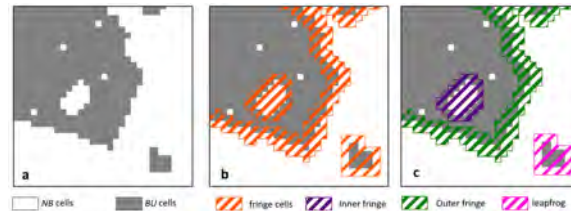
3.1 Recognizing the urban fringe

We recognize the urban fringe by examining, annually, the continuity of built-up (*BU*) and non-built (*NB*) areas. Based on the average distance between buildings in the city, we perform this examination at a resolution of 50m and consider a 50x50m (vector) cell as *BU* if at least 5% of its area is covered with buildings. The cell is a part of the continuous *BU* patch if at least 7 of its 8 neighbors in the 3x3 neighborhood are also *BU*. The same rule applies to *NB* cells, for identifying continuous *NB* patches. An urban fringe is the rest of the area. To group adjacent cells of the *BU* or *NB* types into continuous regions we apply a connected-component labeling algorithm with orthogonal and diagonal (8-cell) connectivity.

The fringe areas $F(t_n)$, estimated at the year t_n , can be of 3 types (Figure 1):

- $F(t_n)$ is an inner fringe, denoted as $F_i(t_n)$, if all cells adjacent to it are of the *BU*-type.

- $F(t_n)$ is a leapfrogging development, denoted as $F_l(t_n)$, if all cells adjacent to it are of the NB -type.
- $F(t_n)$ is an outer fringe, denoted as $F_o(t_n)$, otherwise.



■ **Figure 1** Recognition of a fringe area: (a) construction of BU (grey) and NB (white) continuous areas; (b) fringe area; (c) inner fringe, outer fringe, and leapfrog.

3.2 Recognition of development projects

Definition of a development project: Buildings b_1, b_2, \dots, b_k belong to the same construction project $P(t_n)$ that starts in the year t_n , if (1) they are all recognized, for the first time, in the aerial photo of the year t_n , (2) there is no road between any pair of them, (3) there is no NB areas between them, and (4) they share the same land-use - residential, industrial public, other, determined based on the attributes of the IMC building layer. The spatial extent of the project $P(t_n)$ is established as follows:

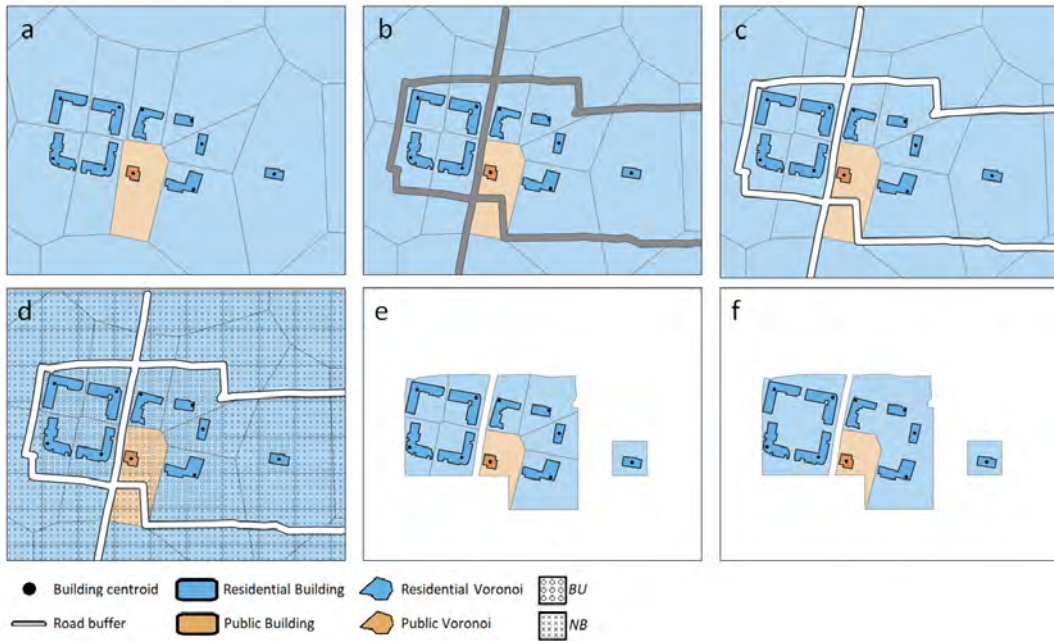
1. Construct Voronoi coverage $V(t_n)$ based on the centroids of all buildings existing at t_n . Assign land use type of the building to its Voronoi polygon (Figure 2a).
2. Construct layer $R(t_n)$ of roads at the year t_n , representing roads as polygons (Figure 2b).
3. Erase road polygons $R(t_n)$ from $V(t_n)$, to obtain corrected Voronoi coverage $V_c(t_n)$ (Figure 2c).
4. Overlay $V_c(t_n)$ and grid G that defines the resolution of our view of the city, currently 50x50 m (Figure 2d).
5. Erase NB polygons (constructed for the fringe assessment) from $V_c(t_n)$ (Figure 2e).
6. Obtain $P(t_n)$ by merging adjacent Voronoi polygons of the same land-use (Figure 2f).

To recognize the changes in the urban patterns between the moments t_{n-1} and t_n we overlap projects that first appeared in the year t_n with the fringe $F(t_{n-1})$. If a certain project $P(t_n)$ overlaps $F_o(t_{n-1})$, then this project is a fringe-expansion. If $P(t_n)$ overlaps $F_i(t_{n-1})$ or is located within the city borders, it is an infilling project. If $P(t_n)$ overlaps $F_l(t_{n-1})$, then it is an old-leapfrog, and if $P(t_n)$ does not overlap $F(t_{n-1})$, it is a new leapfrog (Figure 3).

4 General view of leapfrogging

The amount of new development in Netanya, Haifa, and Safed changed over the 50-year observation period (Figure 4). The construction activities in Haifa and Safed peaked in the early 1990s, while Netanya's period of rapid development was in the early 2000s. The decline in development rate from the year 2000 onwards in all three cities reflects the national trend.

Infilling is the least prevalent form of development in the three cities, accounting, besides the year 2013 in Haifa, for less than 5% of the total construction during the entire period. In the large Haifa and Netanya, fringe projects make up 80-90% of the developed area, while leapfrogging accounts for the remaining 10-20%, except for two spikes in Haifa in 2008 and 2013 with 20-30% of leapfrogging, and Netanya in 2013 with 40% of leapfrogging (Figure 5). In the smaller Safed, the share of leapfrog fluctuates between 25% and 80%, averaging 45%.



■ **Figure 2** Project construction: (a) Voronoi coverage $V(t_n)$ of buildings; (b) Road polygons $R(t_n)$; (c) Road polygons $R(t_n)$ erased from the $V(t_n)$; (d) grid G classified into BU and NB cells; (e) NB polygons erased from the $V_c(t_n)$; (f) $P(t_n)$ is obtained by merging adjacent Voronoi parts of the same land-use.

In the country-wide case of the sprawl of 66 cities between 2013-2018, the average share of leapfrog development is 13% of the total developed area, while the variation of this share is substantial, and the standard deviation of it is 15%. The relationship between population size and the share of leapfrog development is not statistically significant, while if we split the cities into 3 groups – above 100K, 50-100K, and below 50K population, the average shares, by groups, increase from 9% to 14%, and 16%, respectively (Figure 6).

4.1 Adherence to development plans

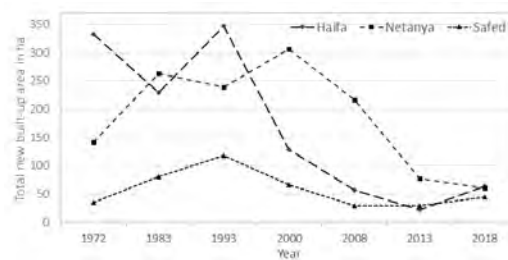
We study adherence to statutory plans by overlaying the layer of the development projects started during the period $[t_n, t_{n+1}]$ and the layer of the development plans for the same period. In this way, we can identify constructions that sprawl beyond the planned areas (usually, to the open lands, agriculture, or forest). Overall, Israeli developers consistently adhere to zoning plans. In Netanya, the overlap is almost 100%, while in Haifa and Safed, 8% to 10% of leapfrog projects violate plan constraints. In the county-wide case, developers also closely adhered to the statutory plans. On average, 94.6% of the projects' area is within the planned border, with 14% of leapfrog projects violating development plans.

4.2 Residential leapfrog project as a seed for future development

We consider leapfrog project $P(t_n)$ erected during period $[t_{n-1}, t_n]$ as an active urban seed if other projects are erected 50 m or less from $P(t_n)$ during the next period $[t_n, t_{n+1}]$. Otherwise, the leapfrog project is passive. Active leapfrogging expresses the system's positive feedback, and its strength can be assessed based on long-term data only. Over 50 years, the share of leapfrog projects in Safed, Netanya, and Haifa that remain passive is 61-68%. Yet, 47-63% of



■ **Figure 3** Leapfrog, infilling, and fringe projects in Afula city between $t = 2013$ and $t = 2018$.



■ **Figure 4** The new built-up area (ha), along the transect, in Haifa, Netanya, and Safed.

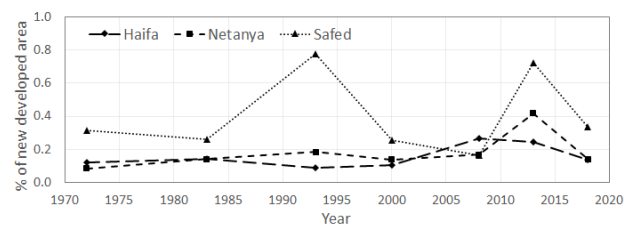
new residential leapfrog projects become active seeds and stimulate additional development. For leapfrog projects of industrial and public land uses, the share of active seeds is much lower and varies between 25-32% and 32-45%, respectively. Residential and industrial seed projects attract projects of the same kind in over 90% of cases across all three cities. Active public projects, on the other hand, exhibit city-dependent attractiveness patterns.

4.3 Urban fringe expansion towards the leapfrog projects

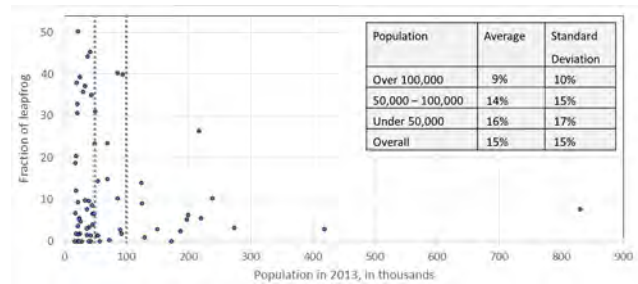
Attracting new constructions, the leapfrog projects become the seeds of unpredictable dynamics of the urban built-up pattern. However, in time, some of these self-organizing areas are absorbed by the regular sprawl of the continuous part of the city. One needs long-term data to estimate the rate of this absorption and to this end, we estimate the percentage of leapfrog projects $P(t_n)$ for which the distance to the nearest project that belongs to the urban fringe becomes zero in time. This assessment demands three sequential observations and we employ it for projects erected until 2008. Estimating this rate, we see that most of the leapfrog projects become absorbed by the city. In Safed, 25% of the projects remain unabsorbed; in Haifa, the share is 31%; and in Netanya, it is 26%.

5 Conclusions

About 13% of the development projects in Israel are leapfrogging and only 14% of these projects, that is, less than 2% of all constructions, violate statutory plans. In time, most of these projects become absorbed by the city, however, before this happens, half of residential and a third of other leapfrog projects serve as seeds for further sprawl. For this reason, leapfrogging often necessitates updates to existing development plans and infractions can be critical for the development trajectory of the city. The importance of leapfrogging as a possible dynamic phenomenon that averts the planned city development trajectory can only be estimated with long-term and high-resolution data on urban dynamics, as above.



■ **Figure 5** The dynamics of the share of leapfrogging development in Haifa, Netanya, and Safed.



■ **Figure 6** The share of the leapfrogging development by cities, depending on their population.

Safed, enveloped by open areas and forests, Haifa with partial constraints, and Netanya, fully surrounded by agricultural lands and other settlements - each city exhibits a unique development pattern, and this pattern is not related to the size of the city. The growth of these cities is mainly defined by historical events, like development peaks in the 90s and early 2000s following mass immigration from the former USSR. We hypothesize that it is this interaction between the external factors and positive plan-violating feedback that makes urban sprawl unpredictable. We plan to explore this unpredictability with an agent-based model of urban growth, whose mechanisms and parameters will be based on the above results.

References

- 1 J. Almagor, I. Benenson, and N. Alfasi. Assessing innovation: Dynamics of high-rise development in an Israeli city. *Environment and Planning B: Urban Analytics and City Science*, 45(2):253–274, 2018.
- 2 D. Broitman and D. Czamanski. Cities in competition, characteristic time, and leapfrogging developers. *Environment and Planning B: Urban Analytics and City Science*, 39(6):1105–1118, 2012.
- 3 Y. Chen, X. Li, X. Liu, and B. Ai. Modeling urban land-use dynamics in a fast developing city using the modified logistic cellular automaton with a patch-based simulation strategy. *International Journal of Geographical Information Science*, 28(2):234–255, 2014.
- 4 J. Portugali. Self-organizing cities. *Futures*, 29(4–5):353–380, 1997.

From Reproducible to Explainable GIScience

Mark Gahegan ✉

School of Computer Science / Centre for eResearch, University of Auckland, New Zealand

Abstract

Communicating deep understanding between humans is key to the effective application and sharing of science, and this is critical in GIScience because much of what we do has practical implications in the modelling and governance of societal and environmental systems. Reproducible and explainable science is needed for public trust, for informed governance, for productivity and for global sustainability [20]. This article summarises some of the more recent research on reproducibility from outside of GIScience, gives practical guidance to current best practice from a GIScience perspective, provides a clearer road-map towards reproducibility and adds in the additional step of explainable GIScience as our final goal.

2012 ACM Subject Classification Information systems → Geographic information systems

Keywords and phrases GIScience, Reproducible, Explainable, discoverable

Digital Object Identifier 10.4230/LIPIcs.GIScience.2023.32

Category Short Paper

1 Introduction

The ‘Reproducibility Crisis’ [2] sent shock waves through both Psychology and Medical Science has changed expectations around how experimental scientists report their research. Apart from the obvious risk of eroding public trust in science if researchers cannot be trusted to behave honestly, reproducibility is critical for two very distinct reasons:

1. For the individual researcher and team, the goal is to discover, access, reuse and build on the work of others, knowing that it can be trusted (efficiency).
2. For the research community as a whole, the goal is to compare new methods, data-sets and theories so we can learn which ones work best, and in what circumstances they can be applied, and to move forward with the best of them (evolution).

Experiments in reproducibility show us that even well-intentioned researchers often fail to provide a complete-enough account of their experiments to allow others to reproduce their results accurately [11]. The bigger issue, then, is not bad actors, but bad practices. The issue has received some good attention of late from the GIScience community [21] including a critical assessment of the reproducibility of GIScience papers published in conference proceedings [14, 15].

Beyond reproducibility is another even more important goal: that of *explainability*. Communicating deep understanding between humans is key to the effective application and sharing of science, and this is critical in GIScience because much of what we do has practical implications in the modelling and governance of societal and environmental systems. Being able to reproduce someone’s research is not enough to ensure it can be successfully repurposed. Repurposing requires that we understand not just the work that was done, but also the situations in which it can be used reliably, and the situations where its underlying assumptions no longer hold. Explainable science is science that can be explored, queried, tested, understood and repurposed, as well as reproduced.



© Mark Gahegan;

licensed under Creative Commons License CC-BY 4.0

12th International Conference on Geographic Information Science (GIScience 2023).

Editors: Roger Beecham, Jed A. Long, Dianna Smith, Qunshan Zhao, and Sarah Wise; Article No. 32; pp. 32:1–32:6

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

2 The journey to Reproducible and Explainable GIScience

We can view the journey towards explainable science as a series of stepping stones, each one taking us a bit closer. A useful starting point is the concise pathway to reproducibility from the *Physiome* journal [16]. Their ideas are expanded in Table 1 below, and a GIScience slant added. Explainability was not included in their text, it has been added in here. Note that there are other definitions in use for some of the terms below, this set has strong traction in the wider sciences.

Replicable Re-running the source code produces a result with reported research. In this case, literally a digital replica of the original experiment produces the same answers. *For example*, source code distributed with a research article (runs and) provides exactly the same results as those documented in the article.

Reproducible When research can, by means of an underlying representation based on domain theory (mathematics, logic or a mix of both), be successfully reproduced in some new system. Source code can be ambiguous and opaque. Logic and mathematics is more precise and often provides more clues as to the semantics. *For example*, a new Geographically Weighted Regression method is successfully re-implemented from a set of equations in a published article. Though not efficient, there are benefits from separately re-implementing methods: it demonstrates that the original description of the method is accurate.

Reusable This requires that the model is well documented, the source code is available and that it is licensed for reuse, so that limitations and appropriate use are clear. Licenses do not remove rights, they add them. In most legal jurisdictions, the absence of any statement about reuse of data or code means that no rights whatsoever are extended. See <https://creativecommons.org/licenses/> for details of which licenses to use. *For example*, code and documentation are managed in a software repository such as github (<https://github.com/>) and the program contains a license statement that enables reuse. The OSGeo Docker image <https://wiki.osgeo.org/wiki/DockerImages> library contains over 70 GIScience applications, ready to install, with documentation and licensing information.

Discoverable Research artifacts can be made discoverable via a metadata description of the content that is accessible to a search engine. As research artifacts have moved online, metadata has been increasingly used to describe the ‘container’ for these artifacts in progressively richer ways. Discovery can be improved by adding in terms that describe the domain and application semantics of artifacts. *For example*, a repository of global landcover maps uses schema(.org) metadata, augmented with the UN’s GlobalLand30 international land-cover categories (<https://www.un-spider.org/links-and-resources/data-sources/land-cover-map-globeland-30-ngcc>) to allow content-based search. State-of-the-art for packaging research artefacts for discoverability and reuse is RO-Crate: <https://www.researchobject.org/ro-crate/>

Validated A method can be considered validated when its predictions under specified conditions match experimental observations. In other words, validation requires that we test a model against real-world observations, not just for consistency within own internal logic or mathematics. Models are typically validated within a range of ‘safe’ operating conditions (such as a scale interval, or between two temperature values). Data-sets can also be validated, or fit-for-purpose. *For example*, a new climate circulation model is validated by several research teams against observed data [18]. We rarely validate in GIScience: we propose new methods, demonstrate that the method works on a test

data-set, but push any comparison to future work. Where a comparison is present, it is often very limited. As a community, we have no real sense which methods are better, nor in what circumstances we should, or should not, use them [8].

Explainable Explanation requires that we can interrogate a model to find out more detail, to clarify our understanding, or to test our assumptions. Such questions could target the data, the code, the theory, the workflow, as well as the more mundane aspects such as the software license or the data-sets used and their reliability and suitability.

For example, a model for political redistricting can reveal to the users relevant details of likely bias and quality issues in underlying data and explain the theory behind the analytical methods employed. No examples exist yet in GIScience.

These six aspects of reproducible science are somewhat entangled. For example, a model can be discoverable without being reusable simply because it does not have an appropriate license information to allow the data or code to be reused.

3 Explainable GIScience – a road-map

The first 5 stepping stones above each add in some useful aspects or hints of explanations, for example by a more provable formal description, or by adding in meta-data. But providing a more complete understanding exactly what has been done in a piece of research, and how, and why, remains challenging. Theory may tell us whether a model is valid, but not how or when to use it; semantics help us to share our ideas and concepts, but does not anchor them into our workflows. Explanations require a complex blend of formal theory, semantics and pragmatics [3, 10, 13] for which there is no conveniently simple packaging.

The challenge in building GIScience explanations is the difficulty in ‘grounding’ geography, that is, to find some scaffolding that is solid enough to build our formal representations upon. The data and concepts used in GIScience are often loaded with complex meaning and abstraction; they can be far removed from physical measurements (though not always). This abstraction also helps explain why it is difficult to come up with laws and theory for GIScience – the data we use are already filtered through so many conceptual lenses that patterns arising from actual measurements are easily lost [9]. So how do we proceed?

3.1 Theory: Connecting Symbolics to Semantics

Symbolic reasoning uses logic, mathematics and other formal theory to represent meaning, with an emphasis on internal consistency and provability, rather than a grounding into semantics. Where the research conducted can be expressed mathematically (e.g. spatial statistical methods), or symbolically using formal logic (some qualitative spatial reasoning methods), then symbolic reasoning provides an excellent grounding into something that is not itself subject to further abstraction – it is foundational. Formal representations seem to have a high currency in the GIScience community, we value the formal grounding of our ideas into symbolic reasoning. (Less so the grounding of our data into suitable ontologies)

But symbolic reasoning by itself is a house of cards. The abstract symbols and functions used are not anchored into any domain semantics, nor into any implementation in a computer program. The reader can often understand them in this way, sometimes with effort, but the process is subjective. Similarly, we can be taught how to translate a symbolic notation from a mathematical form into a computer program. However, translation between a provable abstract representation, the semantics of the domain and the implementation in (say) a program is prone to error. Inconsistencies and translation errors can lead to failures of

reproducibility; misunderstanding and confusion can lead to a failure of explanation. For explanations based only on symbolics there may still be a significant semantic gap and there is no guarantee that the code perfectly implements the equations.

The good news is that symbolics can be tied more closely to both domain semantics and to code, as the following example demonstrate.

LinguaPhylo (LPhy) is a framework to precisely define phylogenetic models (as used for example to understand virus evolution). As the authors state: “*We present a new lightweight and concise model specification language, called ‘LPhy’, that is both human and machine readable. ‘LPhy’ is accompanied by a graphical user interface for building models and simulating data using this new language, as well as for creating natural language narratives describing such models. These narratives can form the basis of manuscript method sections..*” [6]. The code and model examples are here: <https://linguaphylo.github.io/>. LPhy is a programming language designed specifically for a given domain – its operators are those directly used in the domain – rather than abstract types and methods of a traditional programming language. Behind the scenes, and using some clever markup, LPhy creates English language descriptions of the models a user creates. LPhy essentially provides an immutable mapping between the methods that phylogenetics researchers use, the implementation of these methods in code and human-readable descriptions of the resulting workflow.

There is a useful lesson to learn here. Building a bespoke programming language for a large swath of science or geography problems is intractably hard. But if we take a problem that is small enough to have a consistent epistemology, it is possible to create a domain-oriented programming language that is more consistent, reproducible, self-documenting, and explainable, and that makes programming easier. In GIScience, this idea could be used for geostatistical modelling, or to re-engineer tools such as PySal (<https://pysal.org/>) so that they propel GIScience towards reproducible and explainable goals.

Cao et al [4] demonstrate exactly how geographic processes can be represented using geographic and other foundational ontologies. It is these ontologies, then, that need to form the analytical functions in a GIScience LinguaGeo. Of course, geospatial data can also be connected back into ontologies of observations [12] and from there to ontologies of foundational scientific (SI) measurements [17]. The very same anchoring can be used in the representation of variables representing data in the symbolic logic of our methods.

4 What we can do now to encourage reproducibility in GIScience

Replicable Require the publishing of source code and data by all publications that use them.

Encourage journal reviewers to run the code for themselves to establish the truth of the claimed results. Move beyond publishing code to publishing workflows, which also capture additional control flow information.

Reproducible Encourage clear representation of key algorithms in the text of the article. Do not rely solely on source code.

Reusable Insist that all code and data published be made available to other researchers via a permissive license. Ensure that the repositories we use explicitly hold such licensing information (e.g. the OSGeo Docker Repository <https://wiki.osgeo.org/wiki/DockerImages>).

Discoverable Ensure that all data and code are at very least available via a website that is publicly accessible. Use persistent identifiers (DOIs) to ensure longevity. Discoverability is improved by the use of subject-level metadata, so look for repositories that provide this functionality. Even a small amount of metadata is better than none for example see the New York University Spatial Data Repository: <https://geo.nyu.edu/>.

Validated Encourage the validation and comparison of proposed methods, either in the originating article or amongst the wider scholarly community. Use special issues to provide the opportunity for publication of articles that compare methods and that validate published data-sets.

Explainable An open challenge for GIScience is to develop our own LinguaGeo programming language(s) to reduce the gap between code and theory and to automatically generate text descriptions of workflows. In the meantime, we should insist on clear descriptions of methods in text as well as in mathematics or logic. We can also ask for statements that describe any known bias in the data and methods used. For example, if an article examines sentiment analysis in geo-located tweets, what are the socio-demographic biases inherent in these data? Which voices (e.g. ages/genders/ethnicities) are over-represented, and which are not? Where data is being used to train methods, insist that a statement explaining how bias in the data may skew the results obtained. See [19] for more details.

All of these stepping stones, by increasing levels of sophistication, record what was done in precise ways that can survive the process of sharing and so enable researchers to reproduce the findings in a separate computational environment. Some of the responsibility rests with authors, but also some with reviewers and journal editors as well as the scholarly community at large to hold ourselves to a higher standard.

5 The Future: Live and Explainable GIScience?

Perhaps the holy grail of repeatability is a journal article that is itself an executable experiment – that describes an analysis in words, mathematics (or formal logic), semantics and code, but also allows the analysis to be repeated and queried by the reader. A compelling recent example is the *Physiome* journal [16] that encourages authors to submit the analytical models that accompany their more traditional written publications. Physiome evaluates submissions “to determine their reproducibility, reusability, and discoverability. At a minimum, accepted submissions are guaranteed to be in an executable state that reproduces the modelling predictions in an accompanying primary paper, and are archived for permanent access by the community.” The journal uses shared method libraries, process and data ontologies (see [5] for more details), common workflow descriptions and packaged data to deliver on its ambitious claims. It is the culmination of many years of collaborative research within a segment of the bioengineering community. A more general solution to this problem that also maintains dynamic links to changing data (thus updating a publication in real time as new data becomes available) is provided by [7].

6 Conclusion

A lot has been said about whether GIScience is in fact a science [22]. The pros and cons of this argument often revolve around whether GIScience has a unique body of theory that might justify the title. And whilst developing theory is important, it is still only one approach to science [1]. Another is experimentation, and GIScience has much to learn from other experimental sciences in terms of how to report science in ways that are reproducible, understandable by others and that can be easily built upon. Or put another way, GIScience would benefit from acting more like a science in the way we conduct and report our experiments! This article describes the pathway to reproducibility and provides a practical summary of improvements we can collectively make now.

References

- 1 S. T. Arundel and W. Li. *The Evolution of Geospatial Reasoning, Analytics and Modeling In: The Geographic Information Science and Technology Body of Knowledge, John P. Wilson (Ed.)*. UCGIS, 2021.
- 2 M Baker. 1,500 scientists lift the lid on reproducibility. *Nature*, 533:452–454, 2016. doi:10.1038/533452a.
- 3 M. Bunge. How does it work? the search for explanatory mechanisms. *Philosophy of the social sciences*, 34(2):182–210, 2004.
- 4 Y.H. Cao, C.J. Yi, and Y.H. Sheng. Geographic process modeling based on geographic ontology. *Open Geosciences*, 10(1):782–796, 2018.
- 5 M. Clerx, M. Cooling, J. Cooper, A. Garny, K. Moyle, D. Nickerson, P. Nielsen, and H Sorby. Cellml 2.0. *Journal of Integrative Bioinformatics*, 17(2):2020–0021, 2020. doi:10.1515/jib-2020-0021.
- 6 A. J. Drummond, K. Chen, F. K. Mendes, and D. Xie. Linguaphylo: a probabilistic model specification language for reproducible phylogenetic analyses. *bioRxiv*, 2022.08.08.503246, 2022. doi:10.1101/2022.08.08.503246.
- 7 A. Ellerm, B. Adams, M. Gahegan, and L. Trombach. Enabling livepublication. In *IEEE 18th International Conference on e-Science, Salt Lake City, UT, USA*. IEEE Computer Society, 2022. doi:10.1109/eScience55777.2022.00067.
- 8 M. Gahegan. Our gis is too small. *The Canadian Geographer / Le Géographe canadien*, 62:15–26, 2018. doi:10.1111/cag.12434.
- 9 M. F. Goodchild. Commentary: general principles and analytical frameworks in geography and giscience. *Annals of GIS*, 28(1):85–87, 2022. doi:10.1080/19475683.2022.2030943.
- 10 R Harris. *The semantics of science*. A-and-C Black, 2005.
- 11 T. Hothorn and F. Leisch. Case studies in reproducibility. *Briefings in bioinformatics*, 12(3):288–300, 2011.
- 12 K. Janowicz, A. Haller, S. Cox, D. Phuoc, and M. Lefrancois. Sosa: A lightweight ontology for sensors, observations, samples, and actuators. *Journal of Web Semantics*, 2018. doi:10.2139/ssrn.3248499.
- 13 S. Levinson. *Pragmatics*. Cambridge University Press, 1983.
- 14 D. Nüst, C. Granell, B. Hofer, M. Konkol, F.O. Ostermann, R. Sileryte, and V. Cerutti. Reproducible research and giscience: an evaluation using agile conference papers. *PeerJ*, 13(6):p.e 5072, 2018. doi:10.7717/peerj.5072.
- 15 Frank O. Ostermann, Daniel Nüst, Carlos Granell, Barbara Hofer, and Markus Konkol. Reproducible research and giscience: An evaluation using giscience conference papers. In *International Conference Geographic Information Science*, 2020.
- 16 Physiome. The physiome journal. *Physiome*, 1(1):1, 2019.
- 17 H. Rijgersberg, M. Van Assem, and J.L. Top. Ontology of units of measure and related concepts. *Semantic Web*, 4:3–13, 2013.
- 18 F.J. Tapiador, A. Navarro, V. Levizzani, E. García-Ortega, G.J. Huffman, C. Kidd, P.A. Kucera, C.D. Kummerow, H. Masunaga, W.A. Petersen, and R. Roca. Global precipitation measurements for validating climate models. *Atmospheric Research*, 197:1–20, 2017.
- 19 N. Turner Lee, P. Resnick, and G. B. Wednesday. *Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms*. Brookings Institute, 2019.
- 20 R. Vicente-Saez, R. Gustafsson, and C. Martinez-Fuentes. Opening up science for a sustainable world: An expansive normative structure of open science in the digital era. *Science and Public Policy*, 48(6):799–813, 2021. doi:10.1093/scipol/scab049.
- 21 J.P. Wilson, K. Butler, S. Gao, Y. Hu, W. Li, and D.J. Wright. A five-star guide for achieving replicability and reproducibility when working with gis software and algorithms. *Annals of the American Association of Geographers*, 111(5):1311–1317, 2021.
- 22 D. J. Wright, M. F. Goodchild, and J. D. Proctor. Demystifying the persistent ambiguity of gis as “tool” versus “science”. *The Annals of the Association of American Geographers*, 87(2):346–362, 1997.

Uncertainty Quantification in the Road-Level Traffic Risk Prediction by Spatial-Temporal Zero-Inflated Negative Binomial Graph Neural Network(STZINB-GNN)

Xiaowei Gao  

SpaceTimeLab, University College London (UCL), UK

James Haworth  

SpaceTimeLab, University College London (UCL), UK

Dingyi Zhuang  

Department of Urban Studies and Planning, Massachusetts Institute of Technology (MIT), Cambridge, MA, USA

Huanfa Chen  

The Bartlett Centre for Advanced Spatial Analysis, University College London (UCL), UK

Xinke Jiang¹  

School of Computer Science, Peking University (PKU), Beijing, China

Abstract

Urban road-based risk prediction is a crucial yet challenging aspect of research in transportation safety. While most existing studies emphasize accurate prediction, they often overlook the importance of model uncertainty. In this paper, we introduce a novel Spatial-Temporal Zero-Inflated Negative Binomial Graph Neural Network (STZINB-GNN) for road-level traffic risk prediction, with a focus on uncertainty quantification. Our case study, conducted in the Lambeth borough of London, UK, demonstrates the superior performance of our approach in comparison to existing methods. Although the negative binomial distribution may not be the most suitable choice for handling real, non-binary risk levels, our work lays a solid foundation for future research exploring alternative distribution models or techniques. Ultimately, the STZINB-GNN contributes to enhanced transportation safety and data-driven decision-making in urban planning by providing a more accurate and reliable framework for road-level traffic risk prediction and uncertainty quantification.

2012 ACM Subject Classification Information systems → Geographic information systems

Keywords and phrases Traffic Risk Prediction, Uncertainty Quantification, Zero-Inflated Issues, Road Safety

Digital Object Identifier 10.4230/LIPIcs.GIScience.2023.33

Category Short Paper

1 Introduction

In recent years, the field of traffic risk prediction has attracted considerable attention from researchers and policymakers, driven by the need to create resilient urban traffic systems and enhance their reliability in response to mounting challenges such as minimizing urban congestion, improving road safety, and making informed investments in urban infrastructure. Additionally, the zero-inflated nature of accident data, characterized by sparse incidents, poses a substantial challenge to prediction efforts.

¹ Corresponding Author



Deep learning models have emerged as promising tools in this domain, incorporating multivariate spatiotemporal information and utilizing point-processing estimation during training to forecast traffic accidents or risk situations in subsequent time series [2, 12]. For instance, de Medrano et al. [1] proposed a novel SpatioTemporal Neural Network (STNN) framework based on Recurrent Neural Network (RNN) to predict the number of accidents in each region of Madrid, Spain using a 5-hour prediction window. Their results show a more accurate prediction than the traditional linear statistics models as well as machine learning methods. Ren et al. [5] also employed RNN to analyze spatial and temporal patterns of traffic accident frequency and predict grid-level daily risk situations. However, the RNN model is limited by its focus on short-term temporal embedding information and its inability to fully capture the spatial heterogeneity of traffic accidents. Furthermore, Najjar et al. [3] employed Convolutional Neural Networks (CNNs) to combine urban information from satellite imagery and open traffic accident data, mapping city-wide risk situations. Despite this, their approach neglected temporal information and faced limitations due to the quality of street image data.

Recognizing the potential of graph neural networks (GNNs) to account for the natural connection of spatial units, researchers have proposed graph-based models for traffic risk forecasting. Zhang et al. [8] employed a multi-modal approach to jointly consider text data from social media and imagery data from satellites, subsequently mapping grid-level traffic accident frequency using GNNs. Zhou et al. [9] introduced a novel Differential Time-varying Graph Convolution Network (GCN) to dynamically capture traffic variations and inter-subregion correlations, also predicting grid-level traffic risk. After that, they further refined their algorithms to account for hierarchical spatial dependencies, allowing for the mapping of finer grid-level urban traffic risk [10]. While their work addressed the zero-inflation problem of sparse accident data, their models sidestepped this challenge with a grid level and still faced limitations when predicting at the road level, which is a much finer micro-level unit compared.

Despite the valuable foundation provided by predicted average grid-based risk levels, a significant concern in understanding traffic risk prediction is the quantification of uncertainty by considering distribution rather than mean values [11, 4, 12, 7]. Uncertainty is pervasive in urban mobility systems and plays a crucial role in accounting for potential variations in prediction results, which may arise from the aleatoric uncertainty of imbalanced risk data or the epistemic uncertainty of black-box prediction models [2]. Qian et al. [4] explored uncertainty quantification in traffic forecasting by training a graph-based deep learning model to fit aleatoric uncertainty and combining Monte Carlo dropout with Adaptive Weight Averaging re-training methods to estimate epistemic uncertainty. Zhou et al. [11] considered the irregular patterns in human mobility data as aleatoric uncertainty and the average potential variations in predicted results among specific and neighbouring regions as epistemic uncertainty. By recognizing the reducible nature of predicted epistemic uncertainty, they improved prediction performance through a gated mechanism to calibrate the predicted mobility results. Although those two approaches demonstrated the potential of combining variational inference and deep spatial-temporal embedding for predicting various distributions, they did not thoroughly investigate the sparse traffic data scenario. Zhuang et al. [12] and Wang et al. [6] highlighted the importance of considering zero-inflated distribution statistical models for analyzing sparse traffic demand data. These models offer a more accurate spatiotemporal representation of the underlying uncertainty structure suitable for incorporation with deep learning models.

Despite the progress in the field of uncertainty quantification in urban traffic research, the current focus predominantly lies on traffic demands and human mobility. This research mostly utilizes sequence or time-series data and employs coarser resolution prediction approaches, such as grid-based analysis. This also leaves a notable research gap in terms of an investigation into finer resolution models, particularly from a road safety perspective. Road safety prediction involves non-stable and event-based characteristics, which deviate from the usual time-series data analysis. Expanding research in this area could provide a more nuanced understanding of traffic risk prediction and its significant potential for improving urban transportation safety and resilience. Filling this research gap will contribute to the holistic development of urban transportation studies, enhancing not only predictive accuracy but also the applicability of results in practical, real-world scenarios.

Building upon the work of Zhuang et al. [12], this paper presents the Spatial-Temporal Zero-Inflated Negative Binomial Graph Neural Network (STZINB-GNN) model, specifically designed to tackle the existing limitations in traffic risk prediction and uncertainty quantification.

- (a) The zero-inflated negative binomial model is employed to effectively distinguish between non-risk and risk levels across road segments.
- (b) The spatial-temporal Graph Neural Network (ST-GNN) is responsible for learning and fitting the parameters of probabilistic distributions.

To the best of our knowledge, this is the first attempt to merge these two elements for road-level risk estimation. Empirical evidence showcases the enhanced performance of our proposed model when applied to road-daily resolution traffic risk data.

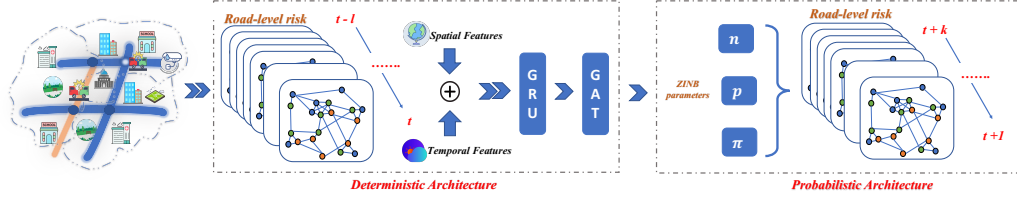
This paper is structured as follows: Section 2 describes the development of the model and provides detailed explanations of its components. Section 3 presents the dataset employed for the case study, outlines the evaluation metrics, and discusses the experimental results. Finally, Section 4 offers conclusions and highlights potential avenues for future research.

2 Methodology

Our objective is to predict future traffic risk and associated confidence intervals for each individual road segment across k forthcoming time intervals, using m roads' risk conditions from previous time windows lasting T days. This is a sequence-to-sequence prediction problem, as illustrated in Figure 1. We construct the road graph, $\mathcal{G} = (V, E)$, where V represents the set of roads, and E denotes the edges connecting these roads. The adjacency matrix $A \in \mathbb{R}^{V \times V}$ indicates the relationships between roads, and $|V| = m \times m$. More specifically, x_{it} signifies the risk level of the i^{th} road segment during the t^{th} time interval, where $i \in V$, $x_{it} \in \mathbb{N}$. Subsequently, $X_t \in \mathbb{N}^{|V| \times T}$ designates the risk conditions for all road segments within the t^{th} time interval, with x_{it} as its component. Our aim is to employ historical records $X_{1:t}$ and spatial-temporal features as input data to estimate the distribution of predicted $X_{t:t+k}$ (i.e., the risk levels for each road over the next k time intervals), thus examining the anticipated value and confidence intervals of the future risk situation.

First, we assume that the inputs follow the ZINB distribution with probability mass function as:

$$f_{ZINB}(x_k; \pi, n, p) = \begin{cases} \pi + (1 - \pi) \binom{x_k + n - 1}{n - 1} (1 - p)^{x_k} p^n & \text{if } x_k = 0 \\ (1 - \pi) \binom{x_k + n - 1}{n - 1} (1 - p)^{x_k} p^n & \text{if } x_k > 0 \end{cases}, \quad (1)$$



■ **Figure 1** Framework of STZINB-GNN model.

where π is the inflation of zeros, n and p determine the number of successes and the probability of a single failure respectively. x_k denotes a road's traffic risk level at one time. All three parameters π, n, p are learned by spatial-temporal GNNs (STGNN), where the temporal encoder applies a Gated Recurrent Unit (GRU) and then the spatial encoder applies Graph Attention Network (GAT):

$$n_{t+1:t+k}, p_{t+1:t+k}, \pi_{t+1:t+k} = STGNN(X_{1:t}; F_{1:t}; A) = GAT(GRU(X_{1:t}; F_{1:t}); A). \quad (2)$$

Here, F_t represents the spatiotemporal features of roads on the t^{th} day, while X_t corresponds to the risk level observed on the same day. This relationship illustrates the connection between road features and risk levels at specific roads in time.

The log-likelihood of ZINB is composed of the $y = 0$ and $y > 0$ parts, and can be approximated as follows:

$$LL_y = \begin{cases} \log \pi + \log(1 - \pi)p^n & \text{when } y = 0 \\ \log(1 - \pi) + \log \Gamma(n + y) - \log \Gamma(y + 1) \\ - \log \Gamma(n) + n \log p + y \log(1 - \pi) & \text{when } y > 0 \end{cases}, \quad (3)$$

where y is the ground-truth value, Γ is the Gamma function and n, p, π is learned by STGNN. The final negative log-likelihood loss function is given by:

$$NLL_{STZINB} = -LL_{y=0} - LL_{y>0}. \quad (4)$$

3 Result

We evaluated the model's performance using a real-world dataset from Lambeth Borough in London, UK. This dataset comprises 5,659 road segments and a total of 1,335 accidents throughout 2019². We calculated a daily risk level by combining the number of accidents with the severity of each accident. We then identified the nearest road segment and accounted for the spillover effects on first and second-order neighbouring roads [9] to assign each road segment a risk value for each day. Notably, the zero-inflation rate for road-level accident risk in Lambeth Borough is 98.7%, indicating that a significant proportion of the road segments experienced no accidents during the observed period.

The evaluation metrics for assessing the model performance are categorized into four aspects. Traditional accuracy measures, including Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and Root Mean Squared Error (RMSE), evaluate the model's ability to accurately predict risk values. Uncertainty quantification is assessed using

² <https://roadtraffic.dft.gov.uk/downloads>

the Kullback-Leibler Divergence (KLD), which indicates how closely the distribution of the model's output risk values approximates the distribution of the true risk values. Lower values for these above metrics are desirable, as they signify a smaller difference between the predicted and actual risk values as well as distributions.

The true-zero rate (ZR) quantifies the model's capacity to accurately replicate the sparsity observed in the ground-truth data. Additionally, the Hit Rate (HR) is assessed based on information entropy. To compute HR, we first select the top 20% of road segments with the highest predicted risk values and then consider the predicted risk values' information entropy, which is derived from the KLD uncertainty quantification. We calculate the HR by selecting those road segments with lower predicted uncertainty among the top 20% risk roads, where the road information entropy is below the mean value of the entire road's entropy.

Higher ZR and HR suggest better model performance in identifying road segments with no accidents and those with accidents, respectively.

■ **Table 1** Model Results for the Lambeth Borough, London.

Results		Long(14-14)				Short(7-7)			
Metrics	Models	STZINB	STG	STNB	HA	STZINB	STG	STNB	HA
ACC	MAE	0.054	0.118	<u>0.080</u>	0.135	0.077	<u>0.048</u>	0.105	0.134
	MAPE	0.026	0.405	<u>0.036</u>	0.414	0.025	0.443	<u>0.078</u>	0.485
	RMSE	0.119	0.183	<u>0.139</u>	0.211	0.139	<u>0.185</u>	0.172	0.238
Uncertainty	KLD	0.259	0.504	1.558	<u>0.269</u>	<u>0.473</u>	0.522	0.759	0.264
Zero Inflated	ZR	0.641	0.199	<u>0.562</u>	0.520	0.579	0.102	<u>0.518</u>	0.503
Hit Rate	HR20%	0.618	0.267	<u>0.447</u>	0.452	0.575	0.301	0.442	<u>0.443</u>

In the table, bold fonts mean the best values among all the baseline models while the underline means the second-best values. The baseline models used for comparison include Spatial-temporal Graph Neural Network with Gaussian Distribution (STG), Spatial-temporal Graph Neural Network with Negative Binomial Distribution (STNB), and Historical Average (HA). It is evident that our proposed model outperforms the baseline models across all evaluation metrics, with the exception of KLD in short-term prediction scenarios, where it ranks second. This demonstrates the effectiveness of our model in capturing the skewed data distribution through its zero-inflated components, which leads to more accurate results and improved reliability when approximating the true risk distribution. Notice that both ZR and HR20% metrics, which are important to measure the occurrence of events in practice, have received significant accuracy improvement. This is due to introducing the parameter π in Equation 1, which can effectively learn the sparsity of the data.

Furthermore, our model's capability to reliably predict higher risk values enables us to achieve an accuracy of approximately 61.8% or 57.5% in identifying the exact locations of accidents, which also significantly outperforms the other GNN counterparts. This highlights the potential of our approach to significantly enhance transportation safety and facilitate data-driven decision-making in urban planning.

4 Discussion and Conclusion

In this study, we developed a versatile spatial-temporal Graph Neural Network (GNN) framework for predicting the probabilistic distribution of sparse road traffic risk and quantifying its associated uncertainty. By employing Gated Recurrent Units (GRUs) to capture temporal correlations and Graph Attention Networks (GATs) to model spatial dependencies, we created a comprehensive framework that embeds the spatial and temporal representations of distribution parameters. These parameters are then fused to obtain a distribution for each spatial-temporal data point.

Our case study, based on the urban risk situation of Lambeth borough in the UK, demonstrated that the proposed model consistently delivers more accurate and reliable results compared to other methods. Despite its performance, the model also has certain limitations. When addressing real risk levels that extend beyond binary values, the negative binomial distribution may not be the most suitable choice. Future work could explore alternative distribution models or techniques that better capture the complexity and nuances of real-world risk levels. This would further enhance the model's applicability and predictive capabilities, ultimately contributing to improved transportation safety and data-driven decision-making in urban planning.

References

- 1 Rodrigo de Medrano and José L Aznarte. A new spatio-temporal neural network approach for traffic accident forecasting. *Applied Artificial Intelligence*, 35(10):782–801, 2021.
- 2 Genwang Liu, Haolin Jin, Jiase Li, Xianbiao Hu, and Jian Li. A bayesian deep learning method for freeway incident detection with uncertainty quantification. *Accident Analysis & Prevention*, 176:106796, 2022.
- 3 Alameen Najjar, Shun'ichi Kaneko, and Yoshikazu Miyanaga. Combining satellite imagery and open data to map road safety. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.
- 4 Weizhu Qian, Dalin Zhang, Yan Zhao, Kai Zheng, and James JQ Yu. Uncertainty quantification for traffic forecasting: A unified approach. *arXiv preprint arXiv:2208.05875*, 2022.
- 5 Honglei Ren, You Song, Jingwen Wang, Yucheng Hu, and Jinzhi Lei. A deep learning approach to the citywide traffic accident risk prediction. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 3346–3351. IEEE, 2018.
- 6 Qingyi Wang, Shenhao Wang, Dingyi Zhuang, Haris Koutsopoulos, and Jinhua Zhao. Uncertainty quantification of spatiotemporal travel demand with probabilistic graph neural networks. *arXiv preprint arXiv:2303.04040*, 2023.
- 7 Yuankai Wu, Dingyi Zhuang, Aurelie Labbe, and Lijun Sun. Inductive graph neural networks for spatiotemporal kriging. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 4478–4485, 2021.
- 8 Yang Zhang, Xiangyu Dong, Lanyu Shang, Daniel Zhang, and Dong Wang. A multi-modal graph neural network approach to traffic risk forecasting in smart urban sensing. In *2020 17th Annual IEEE international conference on sensing, communication, and networking (SECON)*, pages 1–9. IEEE, 2020.
- 9 Zhengyang Zhou, Yang Wang, Xike Xie, Lianliang Chen, and Hengchang Liu. Riskoracle: a minute-level citywide traffic accident forecasting framework. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 1258–1265, 2020.
- 10 Zhengyang Zhou, Yang Wang, Xike Xie, Lianliang Chen, and Chaochao Zhu. Foresee urban sparse traffic accidents: A spatiotemporal multi-granularity perspective. *IEEE Transactions on Knowledge and Data Engineering*, 34(8):3786–3799, 2020.
- 11 Zhengyang Zhou, Yang Wang, Xike Xie, Lei Qiao, and Yuantao Li. Stuanet: Understanding uncertainty in spatiotemporal collective human mobility. In *Proceedings of the Web Conference 2021*, pages 1868–1879, 2021.
- 12 Dingyi Zhuang, Shenhao Wang, Haris Koutsopoulos, and Jinhua Zhao. Uncertainty quantification of sparse travel demand prediction with spatial-temporal graph neural networks. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4639–4647, 2022.

Simulating and Validating the Traffic of Blackwall Tunnel Using TfL Jam Cam Data and Simulation of Urban Mobility (SUMO)

Chukun Gao ✉

Centre for Advanced Spatial Analysis, University College London, UK

Abstract

Blackwall Tunnel is one of the most congested roadways in London. By simulating the tunnel and the connecting roads, information can be obtained about the traffic conditions and bottlenecks. In this paper, a model will be created using the Simulation of Urban Mobility (SUMO) tool and traffic flow data gathered from Transport for London (TfL) traffic cameras. The result from the simulation will be compared to the journey time data of Blackwall Tunnel in order to determine the accuracy of simulation.

2012 ACM Subject Classification Information systems → Traffic analysis

Keywords and phrases Traffic simulation, Validation, SUMO, Blackwall Tunnel

Digital Object Identifier 10.4230/LIPIcs.GIScience.2023.34

Category Short Paper

Supplementary Material *Software (Source Code)*: github.com/Chukun-Leo-Gao/Blackwall_Simulation_GIScience archived at `swh:1:dir:0242fb03acf02ee8c6e971fb8a26814719ac1a1a`

Acknowledgements I would like to thank Transport for London for providing journey time data through its Open Data programme. I would also thank Dr Sarah Wise, my PhD supervisor, for her outstanding support during this research.

1 Introduction

Traffic simulation software is a very powerful tool for such requirement. In the last decade, open-source traffic simulation has been developing at a rapid pace, such as Simulation of Urban Mobility (SUMO) [1], an agent-based traffic simulation program developed in 2001 by the German Aerospace Centre. In this case, SUMO will be utilised to construct a model of Blackwall Tunnel and its connecting roads.

Simulation needs to be supported by – or at least validated against – real and accurate traffic data. Fortunately, Transport for London (TfL) provides traffic camera footage that can be accessed via an API [4]. Traffic camera footage can be an incredibly versatile tool for analysis. It can be utilised for recognition of car makes and models [6]. Calculation of traffic flow count from camera footage has also been carried out [5]. Therefore, TfL traffic camera footage will be used to generate traffic data for the simulation in this paper.

2 Case Study

2.1 About Blackwall Tunnel

Blackwall Tunnel is one of the earliest road tunnels under River Thames in London. It was constructed in 1897, initially with two lanes. It was doubled in 1967, and the current Blackwall Tunnel is operating with four lanes in total [7]. The tunnel is currently one of



© Chukun Gao;

licensed under Creative Commons License CC-BY 4.0

12th International Conference on Geographic Information Science (GIScience 2023).

Editors: Roger Beecham, Jed A. Long, Dianna Smith, Qunshan Zhao, and Sarah Wise; Article No. 34; pp. 34:1–34:8

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

the most congested Thames crossings in London, and a supplement, the Silvertown Tunnel, is now under construction [3]. Therefore, trying to understand the current bottleneck of Blackwall Tunnel will be very helpful for future traffic management of Silvertown Tunnel.

2.2 Literature Review

There have been many attempts at simulating and validating traffic flow in multiple scales, ranging from a whole country to a single motorway. Two main methods are employed. Numerical simulation, as name implies, uses numerical traffic models to estimate traffic flow [8] and are usually limited to one road alone [11]. Agent-based traffic simulation is more versatile. SUMO, as an agent-based traffic simulator, has been put in use in many projects, from small scale simulation such as Bologna city centre [10] to scenarios on a grander scale such as the whole of Luxembourg [2]. However, there has been no previous works on traffic validation of a major roadway using agent-based traffic simulation, so this research fills an important gap within the literature.

3 Methodology

3.1 Constructing the model for SUMO

According to TfL, the Blackwall Tunnel and its approach, dubbed “Blackwall Thoroughfare”, extends from Bow Interchange (Junction with A11) to the north to Sun in the Sands Roundabout (Junction with Shooters Hill Road), stretching a total of approximately 6.9km. In order to compare simulated travel time data with real data from TfL, the main road from Bow Interchange to Sun in the Sands Roundabout, along with slip roads and connecting junctions, should be modelled.

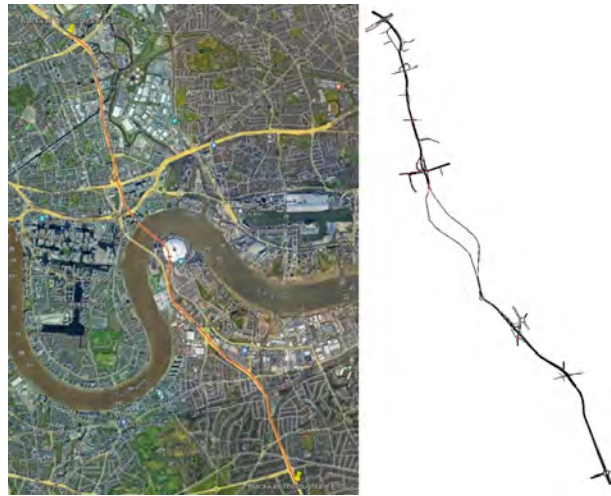
SUMO is supported by a package of powerful tools. In particular, it can read road networks from Open Street Map (OSM), and import the map into NETEDIT program, which is the built-in network editor for SUMO. For this research, Blackwall Thoroughfare and its connecting roads are imported from OSM, and modified to increase realism, including positions and numbers of lanes, shapes of junctions, and placement and timing of traffic lights. The Blackwall Thoroughfare in Google Earth and the same road imported into SUMO NETEDIT can be seen in Figure 1 below.

3.2 Extracting data from TfL Traffic Camera Footage

TfL traffic cameras, or Jam Cams as they are internally known in TfL, provide 10-second footage at 352×288 resolution of live traffic flow. Footage is usually updated once per 5-10 minutes. The Blackwall Thoroughfare is very well covered by Jam Cams, so traffic estimation from the cameras will be relatively accurate.

For this research, all the Jam Cam footages along Blackwall Thoroughfare between 7:30am and 9:30am on all weekdays between 5th and 16th of December for a total of 10 days. As Blackwall Tunnel is the most congested during weekday mornings, the simulation will try to replicate the most stressful condition of the tunnel.

After gathering all the video footages, they are analysed using the “virtual loop” method [13]. Like a traditional traffic induction loop that count cars by detecting magnetic field changes [9], the function of a virtual loop is to count passing vehicles. Firstly, an object detection algorithm, Yolo-v7, is applied to all the Jam Cam footages [12]. Then, a virtual loop, effectively a line drawn across the road in the video footage, is applied, and the number



■ **Figure 1** Blackwall Thoroughfare, Google Earth (left) and SUMO NETEDIT (right).



■ **Figure 2** Typical frame of a TfL Jam Cam footage.

of cars that passes through the loop is then counted. In the image below, the white line represents the virtual loop, and if the coordinate of the bounding box of a car moves across the white line, the vehicle will be counted.

3.3 Simulation Results

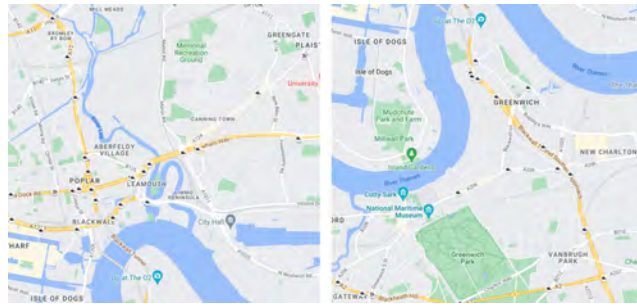
After calculating average traffic flow for every Jam Cam, an Origin-Destination Matrix (OD Matrix) is created for the network, and it serves as the input of traffic simulation in SUMO. The ratio of different vehicle types is shown in the table below. As Blackwall Tunnel has a 13ft (4.0m) height restriction, there are almost no articulated lorries and only a few rigid lorries crossing the tunnel.

■ **Table 1** Ratio of vehicle types.

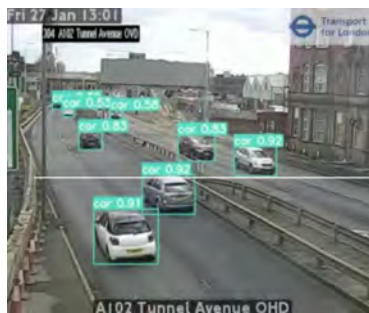
Types	Private Car	Van	Lorry	Articulated Lorry	Motorcycle
Ratio	70%	15%	5%	0%	10%

To further increase the realism of simulation, the departure speeds of the vehicles are randomised. The range of departure speeds of private cars is between 90% and 110% of road speed limit, while lorries are lower (between 70% and 90% of speed limit).

34:4 Traffic Simulation and Validation of Blackwall Tunnel



■ **Figure 3** Jam Cams along Blackwall Thoroughfare (north/south of Thames), each camera symbol is a Jam Cam.



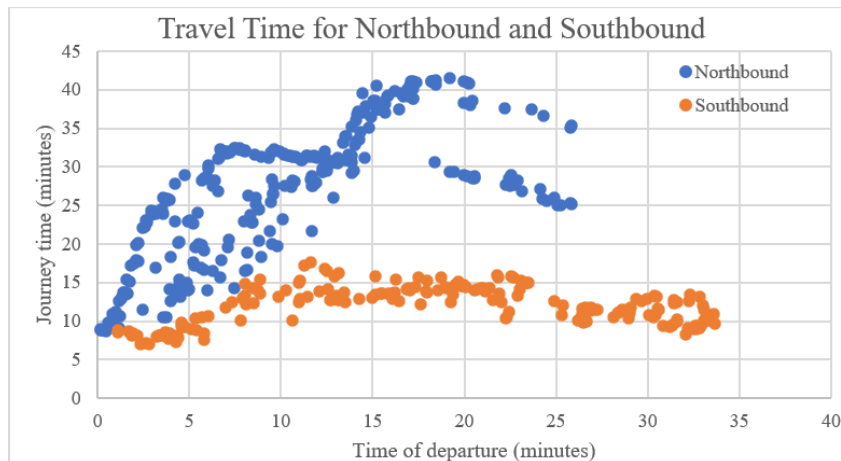
■ **Figure 4** Application of object detection algorithm and virtual loop.

4 Results and Validation

4.1 Simulation Results

As the published travel times are only available for the full length of Blackwall Thoroughfare, only the vehicles travelling from the beginning to the end of Blackwall Thoroughfare is accounted. As TfL journey time data only includes small vehicles, private cars and vans are included in the travel time analysis. In the simulation, 277 cars and vans traverse the Thoroughfare in the northbound direction and 194 traverse in the southbound direction.

In Figure 5, the horizontal axes represent departure time of vehicle after simulation starts, and vertical axes represent time taken to traverse the entirety of Blackwall Thoroughfare. Several phenomena can be observed from the figure. Firstly, the northbound direction traffic takes a lot longer to traverse the Blackwall Thoroughfare, with northbound journey times reaching over 40 minutes, while southbound journey times are within 20 minutes. This is due to a chokepoint at the northbound entrance of Blackwall Tunnel, where three lanes of traffic merges into two. This chokepoint can propagate congestion for more than a mile. The simulation replicates the congestion propagation, which is a positive sign that indicates the simulation can realistically represent real life traffic phenomenon. Secondly, in both northbound and southbound direction, the journey time first remains low, and then gradually increases as simulation progresses. This also happens in real life. During the first several minutes of peak hour (around 7am), the traffic increases drastically, causing congestion.



■ **Figure 5** Journey time with regards to departure time in northbound and southbound direction.

4.2 Data Analysis and Validation

TfL has ceased publishing road journey time data since 19th of May 2021, several days after the easing of the last lockdown. Pre-Covid data, taken from the workdays of the first week of December 2019 (2nd – 6th), will also be provided for comparison. Moreover, TfL provides data for 90th percentile travel time of Blackwall Tunnel, albeit the data was published in 2017. Although it is slightly too outdated for this research, it has been included as a third point of validation.

■ **Table 2** Comparison of simulation data and TfL journey time data.

Travel Time	Unit: Minutes	Average	Standard Deviation	90th percentile
Simulation (December 2022)	Northbound	26.8	8.8	39.3
	Southbound	11.9	2.5	15.0
TfL Data (May 2021)	Northbound	23.7		
	Southbound	8.6		
TfL Data (December 2019)	Northbound	23.9		
	Southbound	10.4		
TfL Data, 90th percentile (2017)	Northbound			52.9
	Southbound			12.9

Apart from travel time, speed is another measurement for traffic simulation. Although TfL does not provide speed data to validate against, it is a better indicator for congestion, and the distribution of speed can help visualise driver behaviour in the simulation. The average speed and 90th percentile for northbound and southbound directions are shown in the table below, and the distributions of northbound and southbound speed are shown in Figure 6.

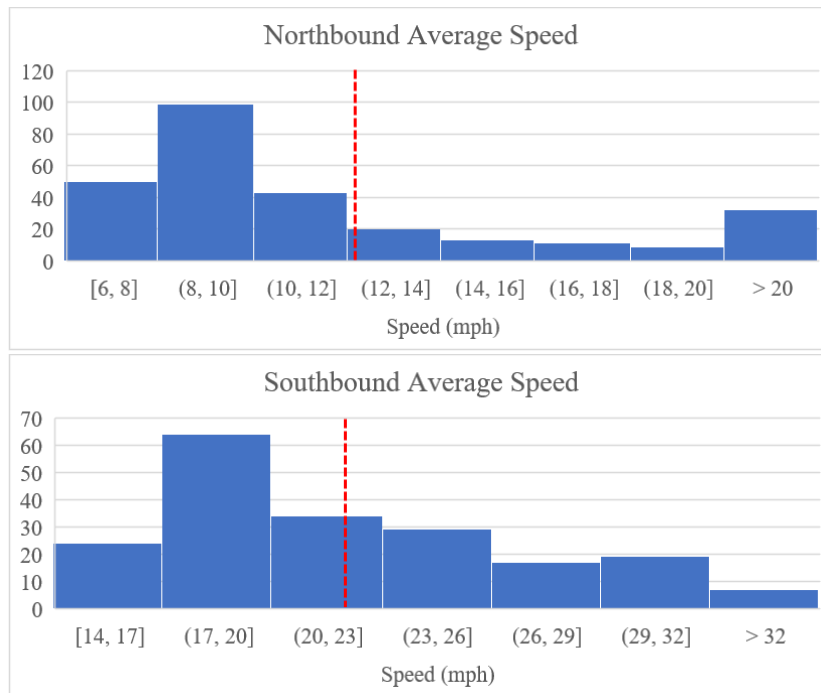
5 Discussion of Data

The simulation corresponds quite well with TfL data from 2019, but less so with data from 2021. The average travel time deviates from 2019 data by 0.4 standard deviation (12%) in northbound direction, and 0.6 standard deviation (13%) in southbound direction. For the 90th percentile, simulated value is 35% lower than the real value in northbound direction, and 14% higher than the real value in southbound direction.

34:6 Traffic Simulation and Validation of Blackwall Tunnel

■ **Table 3** Speed distribution of Blackwall Thoroughfare (lower speed indicates longer journey time).

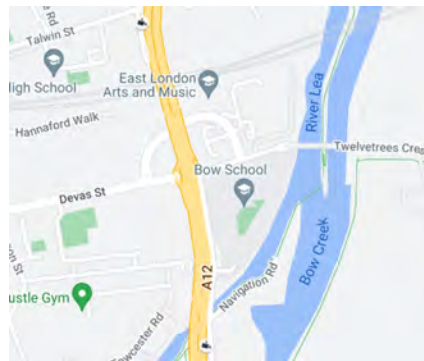
Speed	Unit: mph	Average	Standard Deviation	90th percentile
Simulation (December 2022)	Northbound	12.1	5.9	20.4
	Southbound	22.1	5.0	30.1



■ **Figure 6** Histogram of northbound and southbound average speeds, with red dashed lines denoting the average.

Some insights can be uncovered from the journey time data. First of all, the simulation result is closer to December 2019 than May 2021, showing that the vehicle traffic on the Blackwall Thoroughfare has not yet recovered in May 2021, but has since returned to or even exceeded pre-Covid level in December 2022. Secondly, although the real averages are well within one standard deviation from the simulated values, the overestimates, which are over 10% in both directions, cannot be ignored. This discrepancy can be explained by several factors: TfL Jam Cam data comes in 10-second clips, and can be easily biased. It is possible that by using multiple short clips, traffic flow is overestimated. Also, some important interchanges (e.g., the interchange with Devas Street and Twelvetrees Crescent, see Figure 7) are not covered by any TfL Jam Cams, and traffic flow through the ramps could be overestimated.

Although the 90th percentile data is very old, it can still provide some valuable knowledge. The percentage of overestimate in southbound direction (14%) is similar compared to that of average travel time (13%), meaning the discrepancy might be caused by the same underlying reason. Meanwhile, the simulation severely underestimates the 90th percentile in the northbound direction. It can be caused by accidents in the northbound direction: the “3 into 2” merging at the northbound entrance of Blackwall Tunnel have caused quite a few rear-end collisions in the past year, and accidents can cause further delays. The simulation does not account for accidents, and thus underestimates travel times in extreme cases.



■ **Figure 7** No Jam Cams at the interchange with Devas Street and Twelvetrees Crescent.

Some information could also be gleaned through histograms of average speed. The northbound average speed follows a long tail distribution, with three quarters of cars fall below average speed. Higher speeds correspond to earlier departures, prior to the formation of chokepoint at the northbound entrance of Blackwall Tunnel. However, in the southbound direction, less than 60% of cars fall below average speed, indicating that there is hardly any “early start advantage” on the southbound direction, as there are no chokepoints. This discrepancy is already visible in Figure 5, but Figures 6 and 7 makes it much clearer.

6 Conclusion

In this paper, the Simulation of Urban Mobility (SUMO) tool is used to model the Blackwall Thoroughfare, and the results correspond reasonably well to the data provided by Transport for London (TfL). The simulation slightly overestimates the journey times by approximately 10%, but are well within one standard deviation. This is likely due to the short length of Jam Cam videos and improper Jam Cam coverage. One key gist of this paper is that a large amount of data needed for an accurate agent-based simulation. Without a full picture of the Blackwall Thoroughfare, the simulation will inevitably deviate from reality.

In the future, the author plans to obtain longer videos and videos from junctions without Jam Cam coverage. The footage will be used to determine the traffic within each junction, calculate speed and acceleration distribution, and observe driving behaviour such as lane changes. By incorporating these elements into the simulation, each car agent will become more heterogeneous, thus better mimicking behaviour of real-life drivers. It is hoped that increased heterogeneity will result in a more realistic simulation result.

References

- 1 Michael Behrisch, Laura Bieker, Jakob Erdmann, and Daniel Krajzewicz. Sumo – simulation of urban mobility, 2011.
- 2 Lara Codeca, Raphael Frank, Sebastien Faye, and Thomas Engel. Luxembourg SUMO traffic (LuST) scenario: Traffic demand evaluation. *IEEE Intelligent Transportation Systems Magazine*, 9(2):52–63, 2017. Conference Name: IEEE Intelligent Transportation Systems Magazine. doi:10.1109/MITS.2017.2666585.
- 3 Transport for London. Syndication developer guidelines – transport for london data service, 2012.
- 4 Transport for London. River crossings: Silvertown tunnel – supporting technical documentation, 2014.

- 5 Huan Min Gan, Senaka Fernando, and Miguel Molina-Solana. Scalable object detection pipeline for traffic cameras: Application to tfl JamCams. *Expert Systems with Applications*, 182:115154, 2021. doi:10.1016/j.eswa.2021.115154.
- 6 Hongsheng He, Zhenzhou Shao, and Jindong Tan. Recognition of car makes and models from a single traffic-camera image. *IEEE Transactions on Intelligent Transportation Systems*, 16(6):3182–3192, 2015. Conference Name: IEEE Transactions on Intelligent Transportation Systems. doi:10.1109/TITS.2015.2437998.
- 7 Jasper Kell, Gordon Ridley, and GLC. Blackwall tunnel duplication. *Proceedings of the Institution of Civil Engineers*, 35(2):253–274, 1966.
- 8 Maria Kontorinaki, Anastasia Spiliopoulou, Claudio Roncoli, and Markos Papageorgiou. First-order traffic flow models incorporating capacity drop: Overview and real-data validation. *Transportation Research Part B: Methodological*, 106:52–75, 2017. doi:10.1016/j.trb.2017.10.014.
- 9 Changle Li, Wenwei Yue, Guoqiang Mao, and Zhigang Xu. Congestion propagation based bottleneck identification in urban road networks. *IEEE Transactions on Vehicular Technology*, 69(5):4827–4841, 2020. doi:10.1109/TVT.2020.2973404.
- 10 Pablo Alvarez Lopez, Michael Behrisch, Laura Bieker-Walz, Jakob Erdmann, Yun-Pang Flötteröd, Robert Hilbrich, Leonhard Lücken, Johannes Rummel, Peter Wagner, and Evamarie Wiessner. Microscopic traffic simulation using SUMO. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 2575–2582, 2018. ISSN: 2153-0017. doi:10.1109/ITSC.2018.8569938.
- 11 Tomer Toledo and Haris N. Koutsopoulos. Statistical validation of traffic simulation models. *Transportation Research Record*, 1876(1):142–150, 2004. Publisher: SAGE Publications Inc. doi:10.3141/1876-15.
- 12 Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. doi:10.48550/arXiv.2207.02696.
- 13 Yingjie Xia, Xingmin Shi, Guanghua Song, Qiaolei Geng, and Yuncai Liu. Towards improving quality of video-based vehicle counting method for traffic flow estimation. *Signal Processing*, 120:672–681, 2016. doi:10.1016/j.sigpro.2014.10.035.

Building-Level Comparison of Microsoft and Google Open Building Footprints Datasets

Jack Joseph Gonzales   

Geospatial Science and Human Security Division, Oak Ridge National Laboratory, TN, USA

Abstract

Large-scale datasets of building footprints are a crucial source of information for a variety of efforts. In 2023, the general public benefits from open access to multiple sources of building footprints at the country scale or larger, such as those produced by Microsoft and Google. However, none of the available datasets have attained complete global coverage, and researchers and analysts may need to combine multiple sources to assemble a complete set of building footprints for their area of interest or choose between overlapping sources, requiring an understanding of the differences between different building sources. This paper presents a method to closely examine the quality of different building footprint sources by matching corresponding buildings across datasets, using building footprints in Ethiopia published by Microsoft and Google as an example set.

2012 ACM Subject Classification Computing methodologies

Keywords and phrases Open data, Building footprints, Data comparison

Digital Object Identifier 10.4230/LIPIcs.GIScience.2023.35

Category Short Paper

Acknowledgements The author gives thanks to Daniel Adams and Jessica Moehl for their thoughtful review and advice.

1 Introduction

Among many data resources characterizing the built environment, building footprints have proven to be extremely useful for a wide variety of purposes, from general public use mapping services like OpenStreetMap, to population modeling efforts such as WorldPop and LandScan [10, 1, 11]. At large scale, these building footprints are typically derived from satellite imagery via automated machine learning models, e.g. [14, 13, 12, 8, 5], or using volunteers to manually map out building footprints as in the case of OpenStreetMap [11].

Microsoft and Google have both released expansive datasets of building footprints for use by the general public, providing researchers and analysts with massive datasets covering multiple continents and growing. In addition to their 1.2 billion building dataset covering Europe, much of the Americas, Africa, and Asia, Microsoft has released several independent country-scale datasets, such as the 2018 dataset for the United States [8]. The Google Open Buildings dataset began with a near-complete mapping of buildings in Africa, and has since expanded to parts of Asia and the Americas to include 1.8 billion buildings [5]. Both

© Notice: This manuscript has been authored by UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the US Department of Energy (DOE). The US government retains and the publisher, by accepting the article for publication, acknowledges that the US government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for US government purposes. DOE will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (<http://energy.gov/downloads/doe-public-access-plan>). This material is based upon the work supported by the U.S. Department of Energy under contract no. DE-AC05-00OR22725.



© Jack Joseph Gonzales;
licensed under Creative Commons License CC-BY 4.0

12th International Conference on Geographic Information Science (GIScience 2023).

Editors: Roger Beecham, Jed A. Long, Dianna Smith, Qunshan Zhao, and Sarah Wise; Article No. 35; pp. 35:1–35:6

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

35:2 Comparing Microsoft and Google Open Buildings

Microsoft and Google identify buildings using convolutional neural network-based semantic segmentation models to classify pixels in high-resolution satellite imagery as building or non-building, and then generate building footprint polygons from the positively classified pixels [8, 5, 12].

Other large-scale datasets exist as well, such as EUBUCCO v0.1, which aggregates and harmonizes data from 50 sources to build a dataset of over 200 million buildings for the European Union [9]. OpenStreetMap utilizes a vast number of volunteer analysts to manually map out buildings, providing a good alternative to machine learning-based datasets, albeit very labor intensive to develop and ensure quality, and often lacking in completeness [2, 15, 3, 6].

While all these datasets provide an excellent data resource, they vary in quality and completeness, sometimes requiring multiple sources to be used to completely cover an area of interest. In order to effectively use and integrate data from different sources, effort must be made to understand and account for systemic differences between building footprints from each source. This study presents a framework for comparing one dataset against the other based on matching building footprints from Microsoft and Google.

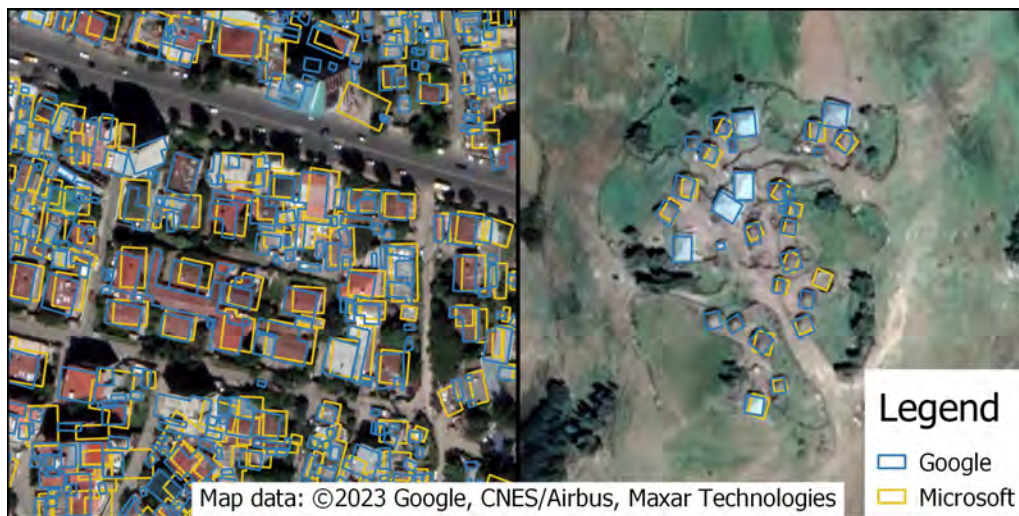
2 Methods

Study area

Two small areas of interest (AOIs) were selected: one from a densely built urban area and another from a low-density rural area. The urban AOI is in the eastern part of Addis Ababa, Ethiopia's capital city, covering roughly 108 hectares and including a good representation of building types found throughout the city. The rural AOI is located in the Amhara region, about 175 kilometers northeast of Addis Ababa, and is dominated by agricultural land with small villages and clusters of buildings scattered about. Examples of the settlement patterns in the AOIs can be seen in the imagery in Figure 1. Like many areas in the world, these AOIs are relatively data poor, with little to no data available other than machine-generated datasets. These two contrasting areas were chosen to evaluate the datasets in a variety of conditions, since settlement patterns heavily differ between urbanized and rural areas, placing different demands on building extraction models. Although small, these AOIs provide a good proof of concept in anticipation of larger-scale comparison efforts.

Data

Building footprints data were sourced from Microsoft's Global Building Footprints and Google's Open Buildings datasets. In addition to footprint geometry, Google provides a confidence value with each footprint, along with guidelines on suggested confidence thresholds to achieve 80%, 85%, or 90% precision. This confidence value allows Google to include many more geometries in their data, many of which may be false detections that can be filtered out using the prescribed confidence thresholds, especially in areas where natural building materials are common and buildings can often be confused with rocks and other landscape features [12]. For this study, we only used those geometries that meet the 90% precision confidence threshold. Microsoft does not report confidence values, but reports that their data achieves 94.4% precision in Africa. Microsoft and Google both report roughly 70% recall [8, 5].



■ **Figure 1** Typical settlement patterns and building footprints in the two study areas, with the urban area on the left and the rural area on the right, overlaid on Google Maps satellite imagery [4].

Comparison

This study seeks to compare matching building footprints from both Microsoft and Google. As such, the initial step is to pair each footprint in one dataset to the footprint(s) that represent the same building in the opposite dataset. For each building in one dataset, matches were identified by identifying all footprints in the opposite dataset that overlap by at least 30% of the area of the smaller geometry, using a similar minimum overlap threshold to Fan et al. 2014 [3]. Individual building footprints may have multiple matches, especially in dense urban areas, where the Microsoft and Google models may disagree on where to divide buildings that are adjacent or have complex, disjointed roofs.

Matched buildings were compared based on area differences and the number of matches found in the other dataset. The number of matches describes the semantic similarity of building detection, or the models' agreement on how to divide complex and adjacent buildings, and can be expressed as a ratio of the number of building footprints in one dataset to the number of corresponding footprints in the other. Possible semantic similarity ratios include 1:1 similarity, where a building matches with exactly one footprint footprint in the other dataset, 1:0 if there is no match, 1:n if one building has multiple matches, m:1 if multiple buildings match one building, or m:n, where multiple buildings match with multiple other buildings [3]. In this study, only 1:1 and 1:n similarity ratios were considered, as other ratios demand a more complex analysis beyond the scope of a short paper, but are important to a complete and thorough examination of the differences between these two sources.

Area comparison is straightforward, taking the median area difference of corresponding footprints between the two datasets, as well as the percentage of buildings with a statistically significant difference from their counterpart in the opposite dataset. A threshold of 1.96 deviations from the median was used to identify values significantly different from the median. Median absolute deviation (MAD) was used to quantify data dispersion as it provides a much more intuitive description of data deviation than the traditional standard deviation [7].

In addition to metrics describing matched building footprints, aggregated statistics describing total number of buildings, percentage of buildings with at least one match, and total and average building area were used to further compare datasets and contextualize statistics of matched buildings.

3 Results

Aggregated statistics

Aggregated statistics shown in Table 1 reveal differing trends for the urban and rural study areas. In the urban area, Microsoft produced fewer, larger building footprints with greater total area, while Google produced more, smaller footprints covering less total area. In both datasets, the majority of buildings had at least one matching footprint.

In the rural area, Microsoft produced far fewer footprints than Google, totalling just 58% of the total area of Google. However, less than half of Google's buildings had a match in Microsoft, whereas 72.5% of Microsoft's buildings had a match. In addition, both produced similar sized footprints on average.

■ **Table 1** Aggregated statistics of each sample set in both urban and rural study areas.

Aggregated Statistics				
Dataset	Total buildings	Percent matched	Total area (ha)	Mean building area (m^2)
Microsoft (Urban)	2,628	66.67	35.03	133.28
Google (Urban)	3,194	72.94	21.68	68.86
Microsoft (Rural)	1,942	72.50	7.02	36.13
Google (Rural)	3,094	46.19	12.11	39.13

Matched building statistics

In the urban area, Google buildings tended to be smaller than their matches in the Microsoft dataset, with a very high MAD, and very few buildings with more than one match, while Microsoft buildings had a higher average number of matches. Both datasets contained similar percentages of buildings with an area significantly different from the median difference.

In the rural area both Microsoft and Google had very similar results, with few buildings matching with more than one other, and Microsoft buildings running slightly smaller than their Google counterparts. MAD for both were nearly identical and far lower than in the urban area. Similar to the urban area, Microsoft buildings had a slightly higher percentage of buildings with a significant area difference.

■ **Table 2** Statistics comparing buildings with their matched counterparts in the opposite dataset.

Matched Area Statistics				
Dataset	Median area difference (m^2)	Area Difference MAD (m^2)	Percent significant difference	Mean number of matches
Microsoft (Urban)	15.96	42.13	15.60	1.39
Google (Urban)	-62.47	131.86	10.48	1.06
Microsoft (Rural)	-2.32	6.65	16.48	1.02
Google (Rural)	2.03	6.55	10.63	1.00

4 Discussion and Conclusion

In the rural study area, matched buildings are remarkably similar, however the aggregated statistics show that Google detected far more buildings, and thus greater total building area. Although many of these buildings have no match in the Microsoft dataset, both datasets

report at least 90% precision and roughly 70% recall, indicating that this discrepancy is most likely predominantly due to different imagery dates and new construction, allowing Google to detect buildings that simply did not exist in the imagery used by Microsoft [5, 8]. This is supported by inspection of Google and Bing satellite maps, with Google imagery appearing to be more recent.

In the urban study area, matched area differences in both datasets show large dispersion, likely due to difficulty in matching the correct buildings with one another. Correctly matching buildings becomes very difficult where imagery is misaligned or models disagree on where to divide and separate buildings. This can be seen on the left side of Figure 1, where overlapping footprints are often very different, as opposed to the rural area on the left where they are very similar. Microsoft tends to generate larger footprints that may encapsulate multiple buildings under a single footprint, while Google tends to break buildings up into smaller polygons, potentially dividing a single complex building into multiple parts. This led Microsoft to generate a larger total building area with fewer buildings, which can be seen Table 1. This discrepancy in polygonization also leads to poor matching results, as small Google footprints may match with a large Microsoft footprint that may completely envelope several Google buildings, leading to the large area difference and dispersion shown in Table 2.

Conclusions

By examining individual building footprints, one can gain a much more in depth understanding of the differences between two data sources that both seek to describe building footprints. This study demonstrates a framework for evaluating differences between two similar sets of polygons, which is crucial for integrating data from multiple sources. It is important to note that neither of these datasets can be considered absolute truth, and rather than determine accuracy, this workflow is designed to characterize differences to assist analysts in integrating or choosing between multiple available data sources. Analysis shows that in the rural area, the Microsoft and Google datasets are very similar where they are able to detect the same buildings, but it is likely that differences in imagery dates result in Google containing additional recently constructed buildings [5, 8]. Differences in the urban area are not likely due to imagery differences, but rather how the models define and separate buildings, as well as difficulty in matching footprints in dense urban areas.



This paper shows an effective method for comparing buildings datasets based on matched footprints in less dense areas, but a more refined matching strategy is needed for an appropriate building-level comparison in highly dense urban areas with complex building patterns. Goals for future work include further development and improvements on the building matching strategy, scaling to larger areas such as regions or countries, and incorporating other building morphology characteristics in addition to area to gain a better understanding of how these different sources characterize the same buildings.

References



- 1 Gianluca Boo, Edith Darin, Douglas R Leasure, Claire A Dooley, Heather R Chamberlain, Attila N Lázár, Kevin Tschirhart, Cyrus Sinai, Nicole A Hoff, Trevon Fuller, et al. High-resolution population estimation using household survey data and building footprints. *Nature communications*, 13(1):1330, 2022.
- 2 Maria Antonia Brovelli and Giorgio Zamboni. A new method for the assessment of spatial accuracy and completeness of openstreetmap building footprints. *ISPRS International Journal of Geo-Information*, 7(8):289, 2018.

- 3 Hongchao Fan, Alexander Zipf, Qing Fu, and Pascal Neis. Quality assessment for building footprints data on openstreetmap. *International Journal of Geographical Information Science*, 28(4):700–719, 2014. doi:10.1080/13658816.2013.867495.
- 4 Google. Google maps, 2023. URL: <https://google.com/maps>.
- 5 Google. Google open buildings, 2023. URL: <https://sites.research.google/open-buildings/>.
- 6 Robert Hecht, Carola Kunze, and Stefan Hahmann. Measuring completeness of building footprints in openstreetmap over space and time. *ISPRS International Journal of Geo-Information*, 2(4):1066–1091, 2013.
- 7 Christophe Leys, Christophe Ley, Olivier Klein, Philippe Bernard, and Laurent Licata. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology*, 49(4):764–766, 2013. doi:10.1016/j.jesp.2013.03.013.
- 8 Microsoft. Microsoft global ml footprints, 2023. URL: <https://github.com/microsoft/GlobalMLBuildingFootprints>.
- 9 Nikola Milojevic-Dupont, Felix Wagner, Florian Nachtigall, Jiawei Hu, Geza Boi Brüser, Marius Zumwald, Filip Biljecki, Niko Heeren, Lynn H. Kaack, Peter-Paul Pichler, and Felix Creutzig. EUBUCCO v0.1: European building stock characteristics in a common and open database for 200+ million individual buildings. *Scientific Data*, 10(1):147, March 2023. doi:10.1038/s41597-023-02040-2.
- 10 J. J. Moehl, E. M. Weber, and J. J. McKee. A vector analytical framework for population modeling. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLVI-4/W2-2021:103–108, 2021. doi:10.5194/isprs-archives-XLVI-4-W2-2021-103-2021.
- 11 OpenStreetMap contributors. Open Street Map retrieved from <https://planet.osm.org> . <https://www.openstreetmap.org>, 2023.
- 12 Wojciech Sirko, Sergii Kashubin, Marvin Ritter, Abigail Annkah, Yasser Salah Eddine Bouchareb, Yann Dauphin, Daniel Keysers, Maxim Neumann, Moustapha Cisse, and John Quinn. Continental-Scale Building Detection from High Resolution Satellite Imagery, July 2021. Number: arXiv:2107.12283 arXiv:2107.12283 [cs]. URL: <http://arxiv.org/abs/2107.12283>.
- 13 Benjamin Swan, Melanie Laverdiere, H. Lexie Yang, and Amy Rose. Iterative self-organizing SCENE-LEVEL sampling (ISOSCELES) for large-scale building extraction. *GIScience & Remote Sensing*, 59(1):1–16, December 2022. doi:10.1080/15481603.2021.2006433.
- 14 Hsiuhan Lexie Yang, Jiangye Yuan, Dalton Lunga, Melanie Laverdiere, Amy Rose, and Budhendra Bhaduri. Building Extraction at Scale Using Convolutional Neural Network: Mapping of the United States. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 11(8):2600–2614, August 2018. doi:10.1109/JSTARS.2018.2835377.
- 15 Qi Zhou, Yuheng Zhang, Ke Chang, and Maria Antonia Brovelli. Assessing osm building completeness for almost 13,000 cities globally. *International Journal of Digital Earth*, 15(1):2400–2421, 2022.



Characterizing Urban Expansion Processes Using Dynamic Spatial Models – a European Application

Alex Hagen-Zanker  

School of Sustainability, Civil and Environmental Engineering, University of Surrey, UK

Jingyan Yu  

Institute of Geography and Sustainability (IGD), Faculty of Geosciences and Environment, University of Lausanne, Switzerland

Naratip Santitissadeekorn  

Department of Mathematics and Physics, University of Surrey, UK

Susan Hughes  

School of Sustainability, Civil and Environmental Engineering, University of Surrey, UK

Abstract

Characterisation of the urban expansion processes using time series of binary urban/non-urban land cover data is complex due to the need to account for the initial configuration and the rate of urban expansion over the analysed period. Failure to account for these factors makes the interpretation of landscape metrics for compactness, fragmentation, or clumpiness problematic and the comparison between geographical areas and time periods contentious. This paper presents an approach for characterisation using spatio-dynamic modelling which is data-centred using a process based model, Bayesian optimization, cluster identification, and maximum likelihood classification. An application of the approach across 652 functional urban areas in Europe (1975-2014) demonstrates the consistency of the approach and its ability to identify spatial and temporal trends in urban expansion processes.

2012 ACM Subject Classification Applied computing → Environmental sciences

Keywords and phrases Urban expansion, morphology, spatio-temporal dynamics, simulation, compactness

Digital Object Identifier 10.4230/LIPIcs.GIScience.2023.36

Category Short Paper

Funding NERC(UKRI) Landscape Decisions project NE/T004150/1.

1 Introduction

Urban expansion along with climate change is one of the major global challenges, affecting all pillars of sustainable development. Past processes of urban expansion are often characterised in terms of composition, for example by the rate of growth of built-up areas. However, it is also of relevance to understand the spatial structure, i.e. the spatial configuration and its process of change. In particular the compactness of urban areas is consequential as it affects the quality of both the natural (e.g. fragmentation of habitats) and urban (e.g. transport demand, walkability) environment.

Commonly, as in this paper, the source data for analysis of urban expansion is multi-temporal raster data classified into binary urban/non-urban classes. The methods that are widely used for the characterization of urban configuration include landscape metrics that were largely developed and applied in the field of landscape ecology. These metrics include the dispersion index, clumpiness index, fractal dimension and compactness index. These metrics can characterize temporal change when applied cross-sectionally for multiple moments in time. Few metrics exist that take a longitudinal perspective and characterize changes over



© Alex Hagen-Zanker, Jingyan Yu, Naratip Santitissadeekorn, and Susan Hughes; licensed under Creative Commons License CC-BY 4.0

12th International Conference on Geographic Information Science (GIScience 2023).

Editors: Roger Beecham, Jed A. Long, Dianna Smith, Qunshan Zhao, and Sarah Wise; Article No. 36; pp. 36:1–36:6

Leibniz International Proceedings in Informatics



LIPIC Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

time. Notable exceptions are the Landscape Expansion Index[3], which measures to what extent new urban land is adjacent to existing urban land and the classification of change events as infill, edge expansion, or leapfrogging[8].

The suitability of the landscape metrics to describe urban expansion *processes* is limited: the same observed changes in landscape metrics may be the result of different processes; Furthermore, the same process will have different effects on landscape metrics dependent on the initial configuration as well as the duration over which the processes are active. This paper investigates an alternative approach to characterizing urban expansion processes. The rationale is to characterize the urban expansion that occurs over a given period by the simulation model that best describes the observed changes. The initial configuration is exogenous to the model, as is the total area of expansion. Hence, the model – and classification – are exclusively about the change in urban configuration. The urban expansion model used is the recent model by Yu et al. [6] as is the clustering of parameter sets into four growth modes ranging from compact to dispersed [7]. This current paper extends this work by applying the classification method to 652 functional urban areas (FUAs) in OECD countries within Europe over the periods 1975-1990, 1990-2000 and 2000-2014. For a sample of FUAs the characterization will be compared to the well-established metrics of fractal dimension (FD)[2] and dispersion index (DI)[5].

2 Methods and data

The model is a Constrained Cellular Automata urban expansion model. It is dynamic in the sense that it starts from an initial urban configuration and then steps through time to incrementally allocate new urban land to raster cells. The model takes the total urban land at each moment in time as an exogenous constraint. The model represents complex dynamics as the spatial configuration of existing urban land is the main factor determining the locations where new urban expansion takes place, causing a process of self-organisation. With just four parameters, representing processes of agglomeration and preservation of natural capital it is one of the most concise urban expansion models. The use of the model to characterise urban expansion patterns goes through several stages:

1. The first stage is calibration using a stochastic method based on Markov chain Monte Carlo with approximate Bayesian computation. For each FUA and time period it produces twenty different parameter sets representing the uncertainty of the calibration. Yu et al. [7] estimated the model for ten FUA across Europe and two time periods and thus produced $10 \times 2 \times 20 = 400$ parameter sets.
2. In the second stage the generated parameter sets are applied to a common initial configuration and rate of urban expansion yielding 400 simulated urban configurations.
3. In the third stage all 400 simulated urban configurations are mutually compared and clustered into four groups based on their similarity. The four groups are considered urban expansion modes and were labelled 'compact', 'medium compact', 'medium dispersed' and 'dispersed'.
4. The fourth stage of the classification applies sample parameter sets from each of the urban expansion modes to a single FUA over a given period. A basic maximum likelihood classification takes place based on the urban expansion mode that most closely resembles the observed dynamics.

This paper uses the model and parameter clusters identified before and extends the analysis to the full set of 652 FUAs within European OECD countries. The built-up and functional

urban area data that support the findings of this study are part of the Global Human Settlement Layer (GHSL)[1] [4]. All the models and analyses of this study are implemented in Python as open-source¹.

3 Results and discussion

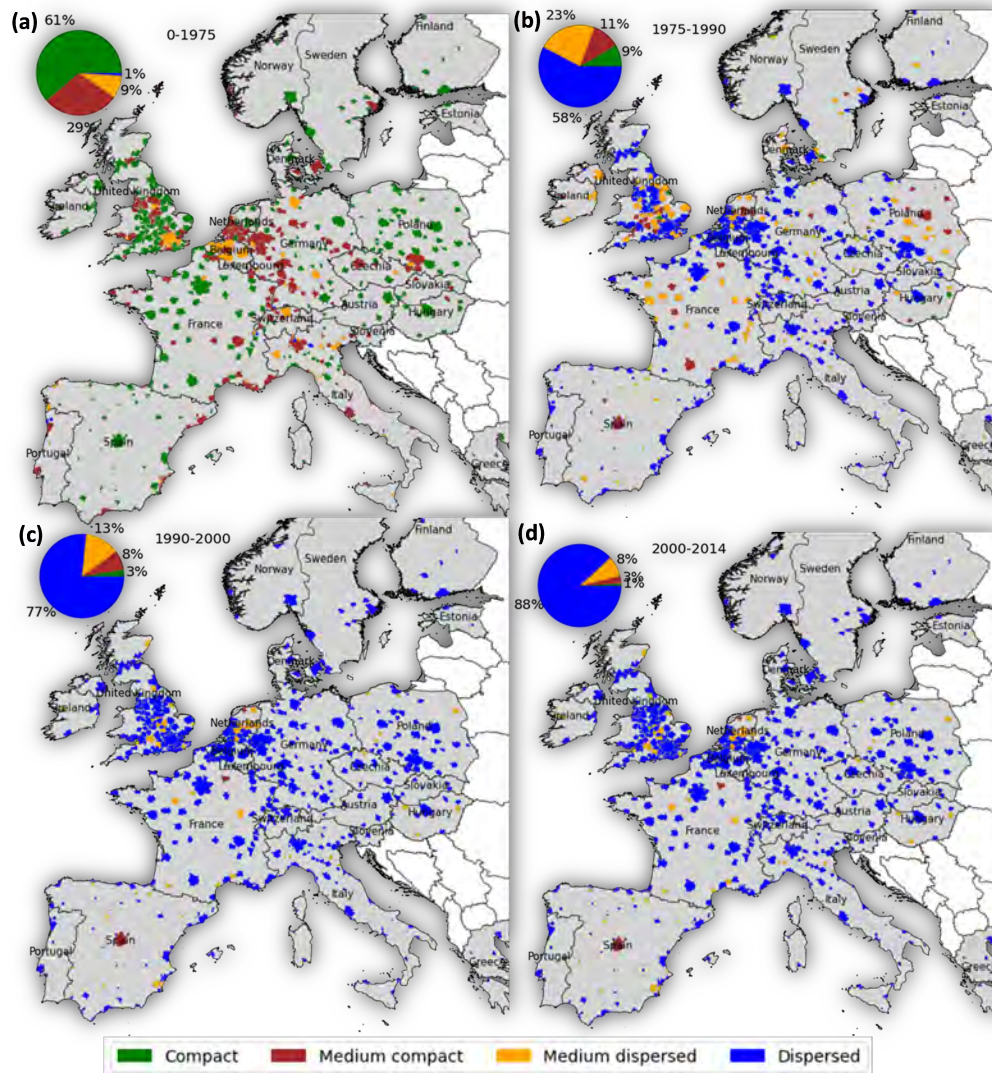
The results, as seen in Fig. 1 present classification of urban expansion processes in Europe over time. The first period is from “0” to 1975, this classification is based on the urban expansion mode that best represents the expansion from urban genesis (a map void of urban land) to 1975. The results indicate that the processes that have historically shaped urban form in Europe could best be described as compact or medium compact. From 1975 onward however, a clear shift is visible and increasingly over time, more FUAs are becoming classified as undergoing dispersed or medium dispersed expansion processes. This does not imply that this shift occurred in 1975, but rather that it occurred sometime *before* 1975. Where in 1975-1990 58% of FUA could be classified as having a dispersed urban expansion process, in 2000-2014 this had increased to 88%. There is also a distinct spatial pattern, of more urban and industrialized areas turning towards a dispersed process of expansion first, and more rural areas following later.

For a sample of four FUAs we show four model realisations of urban expansion patterns (one for each mode), as well as the observed urban expansion pattern (Fig. 2). For each of the resulting maps the corresponding Dispersion Index and Fractal Dimension are also calculated. The results indicate that the four urban expansion modes reflect a variability of modelled expansion patterns that reflects actual variability across time and FUAs. The comparison of compactness metric by urban expansion mode (Fig. 3) shows that for each of the four FUAs individually the results are consistent, i.e. a more dispersed expansion mode is reflected in corresponding values for DI and FD. However between FUAs the results are not comparable: based on the metrics alone it is not possible to predict what expansion mode a FUA belongs to. Efforts to make the metrics more comparable, by considering the relative change of the metric over time, or by considering the relative metric value compared to that of the compact expansion scenario, did not effectively make the results more comparable (Fig. 3.) These results supports the assertion in the introduction that existing landscape metrics are ill-suited to give insight in urban expansion processes when there is variation in initial configuration or rate of expansion.

4 Conclusion

The proposed method for characterising urban expansion processes presents stark spatio-temporal patterns of changing urban expansion processes across Europe in recent decades. The method is complex and computationally intensive, but is more effective than widely used landscape metrics in characterizing urban expansion processes. The reason for this is that the simulation model based approach is inherently dynamic and independent of initial configuration and quantity or rate of expansion. Although specifically aimed at the process of urban expansion, the general framework should be applicable to a wider range of spatial dynamics.

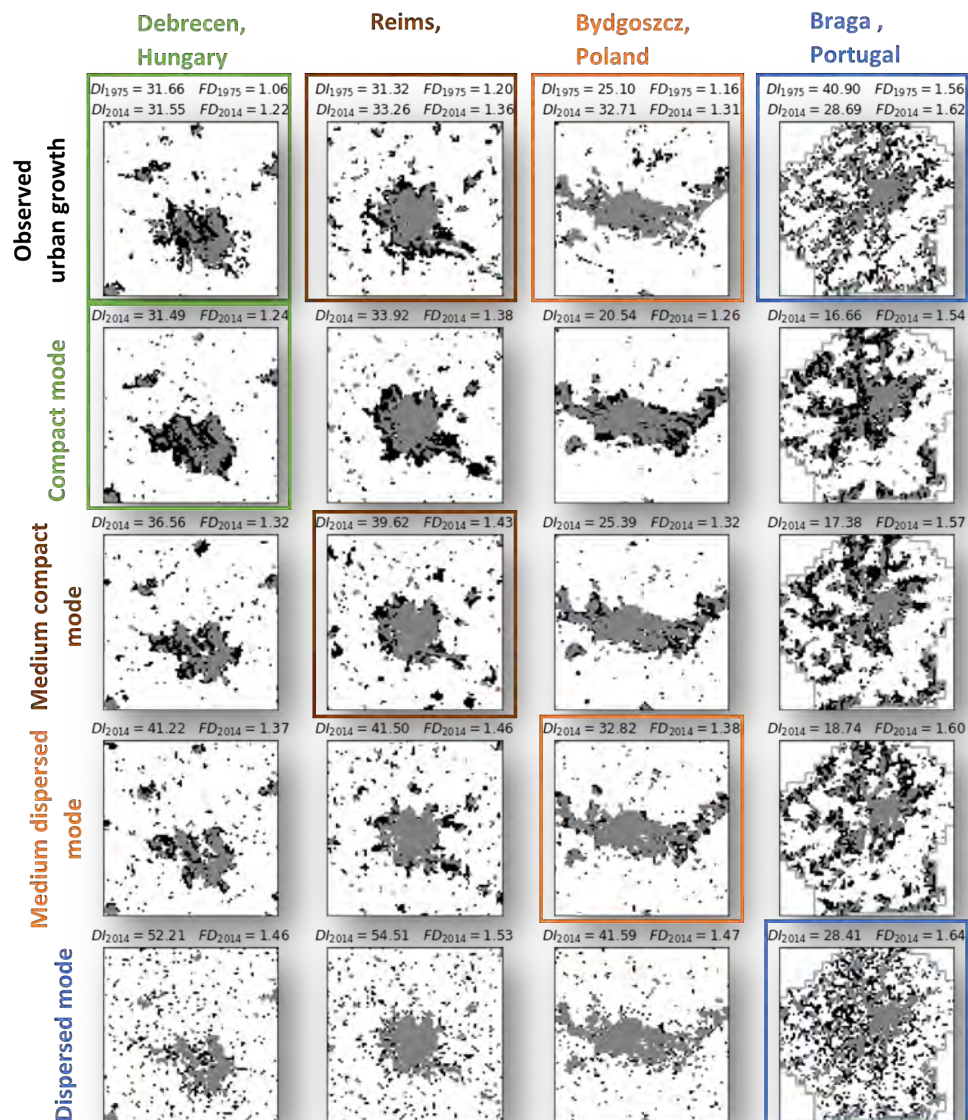
¹ Available here: <https://github.com/JingyanYu/LandUseDecisions>



■ **Figure 1** Classification of urban expansion processes for FUAs in Europe over time.

References

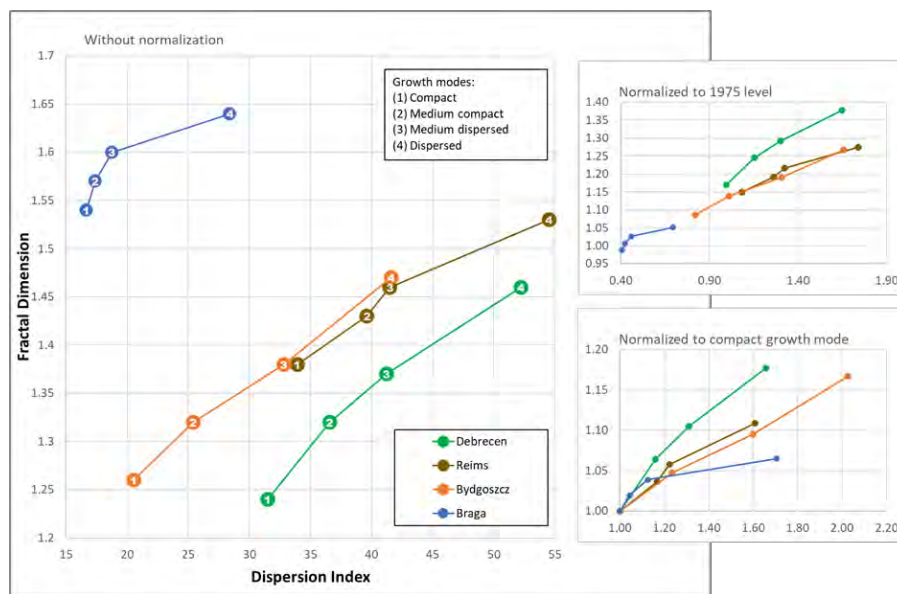
- 1 Christina Corbane, Aneta Florczyk, Martino Pesaresi, Panagiotis Politis, and Vasileios Syrris. GHS-BUILT R2018A - GHS built-up grid, derived from Landsat, multitemporal (1975-1990-2000-2014). *European Commission, Joint Research Centre (JRC)*, 2018. doi:10.2905/jrc-ghs1-10007.
- 2 Atilla R. Imre and Jan Bogaert. The Minkowski-Bouligand dimension and the interior-to-edge ratio of habitats. *Fractals*, 14(01):49–53, 2006. doi:10.1142/S0218348X06003027.
- 3 Xiaoping Liu, Xia Li, Yimin Chen, Zhangzhi Tan, Shaoying Li, and Bin Ai. A new landscape index for quantifying urban expansion using multi-temporal remotely sensed data. *Landscape ecology*, 25:671–682, 2010. doi:10.1007/s10980-010-9454-5.
- 4 Marcello Schiavina, Ana Moreno-Monroy, Luca Maffenini, and Paolo Veneri. GHS-FUA R2019A—GHS functional urban areas, derived from GHS-UCDB R2019A,(2015).



■ **Figure 2** Model realisations for four sample FUA for each growth mode, and observed expansion (1975-2014).

R2019A. edited by Joint Research Centre (JRC) European Commission, 2019. doi:10.2905/347F0337-F2DA-4592-87B3-E25975EC2C95.

- 5 Hannes Taubenböck, Michael Wurm, Christian Geiß, Stefan Dech, and Stefan Siedentop. Urbanization between compactness and dispersion: Designing a spatial model for measuring 2d binary settlement landscape configurations. *International Journal of Digital Earth*, 12(6):679–698, 2019. doi:10.1080/17538947.2018.1474957.
- 6 Jingyan Yu, Alex Hagen-Zanker, Naratip Santitissadeekorn, and Susan Hughes. Calibration of cellular automata urban growth models from urban genesis onwards—a novel application of Markov chain Monte Carlo approximate bayesian computation. *Computers, Environment and Urban Systems*, 90:101689, 2021. doi:10.1016/j.compenvurbsys.2021.101689.



■ **Figure 3** Comparison with widely used metrics of urban form.

- 7 Jingyan Yu, Alex Hagen-Zanker, Naratip Santitissadeekorn, and Susan Hughes. A data-driven framework to manage uncertainty due to limited transferability in urban growth models. *Computers, Environment and Urban Systems*, 98:101892, 2022. doi:10.1016/j.compenvurbsys.2022.101892.
- 8 Hui Zeng, Daniel Z. Sui, and Shujuan Li. Linking urban field theory with GIS and remote sensing to detect signatures of rapid urbanization on the landscape: Toward a new approach for characterizing urban sprawl. *Urban Geography*, 26(5):410–434, 2005. doi:10.2747/0272-3638.26.5.410.

Understanding the Spatial Complexity in Landscape Narratives Through Qualitative Representation of Space

Erum Haris ¹  

School of Computing, University of Leeds, UK

Anthony G. Cohn  

School of Computing, University of Leeds, UK
The Alan Turing Institute, London, UK

John G. Stell  

School of Computing, University of Leeds, UK

Abstract

Narratives are the richest source of information about the human experience of place. They represent events and movement, both physical and conceptual, within time and space. Existing techniques in geographical text analysis usually incorporate named places with coordinate information. This is a serious limitation because many textual references to geography are ambiguous, non-specific, or relative. It is imperative but hard for a geographic information system to capture a text's sense of place, an imprecise concept. This work aims to utilize qualitative spatial representation and natural language processing to allow representations of all three characteristics of place (location, locale, sense of place) as found in textual sources.

2012 ACM Subject Classification Computing methodologies → Knowledge representation and reasoning

Keywords and phrases Narratives, Qualitative spatial representation, Natural language processing

Digital Object Identifier 10.4230/LIPIcs.GIScience.2023.37

Category Short Paper

Supplementary Material *Dataset:* <https://github.com/UCREL/LakeDistrictCorpus>

Funding The support of the Economic and Social Research Council (ESRC) under grant ES/W003473/1 is gratefully acknowledged.

Acknowledgements We also thank the entire team of the Spatial Narratives project (<https://spacetime Narratives.github.io/>) for their discussions on the CLDW and this work.

1 Introduction

Narratives are a fundamental way of organizing experiences and giving them meaning. Narrative theory privileges time by emphasizing the sequence of events, yet all narratives also imply a material, spatial world. Narratives represent events and movement, both physical and conceptual, within time and space. Most work to date looking at geographies within digitized texts has focused on extracting and mapping well-recognized toponyms i.e. place names with geographic coordinates. However, in practice, people conveniently log and share their narrative experiences in imprecise natural language. They likely recall locations qualitatively [6]. For instance, they would share information like “Lake Gardens is a 10-minute walk east

¹ Corresponding author. **Contributions:** EH wrote the original submission based on work done by AGC; AGC and JGS acquired the funding and commented on the revised version.



of the monorail station” instead of “From latitude, longitude 3.177383, 101.7076 walk to 3.1733, 101.6959” . In such cases, qualitative and approximate spatial statements are more useful than exact location coordinates [5].

The term “landscape narratives” describes the interaction and mutual relationship between story and place. Place has multiple characteristics including: the location of an object or event; the natural and built physical environment that makes up a place, termed its locale; and sense of place, the accumulated events, actions, and memories attributed to a location [1]. In this work, we aim to preserve the narrative structure of text data, and move much further by incorporating geographical features that cannot be easily mapped (“a town”, “the hills”), relative spatialities (“near to”, “20 minutes from”), and senses of place (“picturesque”, “enclosed”). This will be achieved by combining natural language processing (NLP) and qualitative spatial and temporal reasoning (QSTR) to extract, locate, and contextualise imprecise information about places.

This work particularly explores the Corpus of Lake District Writing (CLDW) consists of travel writing and tourist literature from 1622 to 1900 describing the English Lake District. It contains 80 geoparsed texts by famous writers, such as Wordsworth and Coleridge, and works from lesser-known writers and travel guides. The corpus portrays leisure journeys where the aim is to describe the landscape and evoke an emotional response. It offers us the opportunity to assess how human experiences of space are represented in writing. It has already been extensively analysed in geographical information systems (GIS) using quantitative places [3] [13]. Applying our methods to these texts will allow us to develop an enhanced understanding of what semantic, grammatical, and geographical tropes can be discovered in individual texts and entire corpora, allowing us to better understand the senses of place recorded by individual writers and their aggregate grouping. The scope of the details provided in this paper is limited to spatial information part of the larger work.

2 Qualitative Spatial Representation for Textual Data: Background and related work

Spatial representation usually locates objects in a quantitative frame. Natural language, on the other hand, offers an imprecise and vague setting. The emergence of qualitative spatial representation (QSR) provided a systematic description of space in this domain. It defines locations as regions, but without the need for geometric information. It uses a common-sense level of abstraction to represent spatial knowledge in terms of connections or spatial relationships between one region and another region. Hence, spatial relations are one of the fundamental aspects for formal qualitative representation of space. Examples include topological, direction and distance associations [5]. The difficulty lies in analysing textual data to identify connections between regions, so one can perceive spatial representations that go beyond toponyms [12].

Additionally, QSR makes use of the logical properties of relationships between entities, enabling data consisting of entities with qualitative relationships between them to be handled as a network of nodes and labelled links. This provides a computational representation of qualitative data even in cases, such as geographical relationships, where specific details are unknown [10]. In this way, it can push spatial study beyond the limitations imposed by quantitative geographical information. Some notable and relevant studies include the work on the analysis of the 16th century Mexican maps [10] to model complex and diverse spatial information, including social and symbolic conceptions depicted in the maps. The study explores the implications of qualitative spatial reasoning for historical and archaeological

research. Another interesting study [9] adopted this technique along with corpus linguistics and NLP in humanitarian forensic research to analyze social and media releases from official sources to gain an understanding of the death of migrants at the Texas-Mexico border. This illustrates the utility of qualitative spatial representation in Humanitarian GIS. Semantic role labelling [7] draws on natural language semantics for the extraction of qualitative descriptions from text is yet another application of qualitative spatial reasoning. A pertinent study [8] presents mapping of natural language to formal spatial representation using a two-level approach where the first level deals with spatial role labelling and the second level maps these arguments to formal spatial calculi.

3 Proposed Methodology

This work is part of a larger project focusing on the extraction, qualitative representation, analysis and visualization of the CLDW [11]. Each phase of the project forms an independent module. In this context, the proposed work will utilize QSTR to demonstrate how one can extract a network of geographical and temporal entities and relationships combining location, locale, and sense of place from the CLDW. The first stage is the development of an appropriate reasoning mechanism using existing and extended qualitative spatial and temporal calculi which refer to the sets of relationships encoding spatial and temporal semantics with associated inference mechanisms. This stage will primarily use the annotations from the first module which focuses on the corpus linguistics and NLP techniques for named-entity recognition (NER) and related tasks [4]. Spatial relations can be understood from multiple, sometimes conflicting viewpoints. Existing research [2] identifies a “user level” corresponding in our case to the writer’s intended meaning. Two other levels in [2] are called “geometrical” and “computational” which provide respectively an abstract mathematical denotation and an implementation. An alternative abstract level to [2] is to specify meanings of relations by logical formulas. As part of the first stage, we aim to develop an ontology to define various categories of spatial and temporal entities and relations exist at the user level. The identified entities and relations will be used to construct semantic triples. An analysis of the meanings of spatio-temporal relationships in the corpus requires the transformation of these extracted user-level triples to a suitable abstract level, and then to the computational level to connect with geographical visualizations. The overlap between these relationships and existing spatio-temporal calculi will be identified. Calculi will be designed that extend current AI-focused work in QSTR to narratives. This will not only allow the expression of qualitative relationships, but also those which are vague and imprecise.

In the second stage, semantic representations of narratives will be developed as networks with locations, temporal entities and events as network nodes and spatio-temporal relationships as edges, or links between nodes. This will be followed by a network analysis step since patterns within networks will represent more complex relationships. Hence, the overall task is to explore the extent to which existing QSTR reasoning methods are adequate to allow deeper meanings to be extracted from these representations and design new inference methods to allow hidden meanings and consequences to be made explicit.

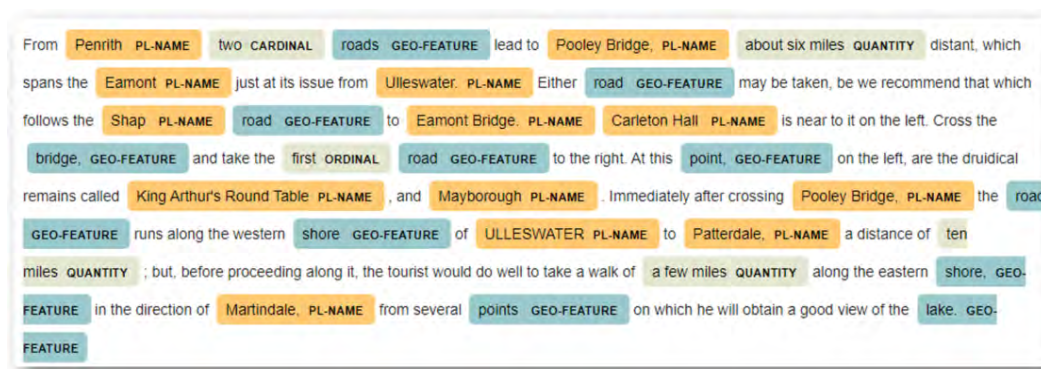
3.1 Representing spatial information using QSR

This section provides details on representing spatial information using QSR. It presents an interpretation of the different types of qualitative relations found in the CLDW as narrative writing. The purpose here is to illustrate the complexity of describing space, which in turn requires a range of inference mechanisms to appropriately represent respective relations.

37:4 Understanding the Spatial Complexity

Hence, it is imperative to provide a non-mathematical description of the mapping from object-entity relation to possible geometric space representation using QSR inference rules. From the given examples, it can be realized that some of the relationships are straightforward, while others being abstract require development of inference mechanisms. A few relations are highly domain-specific and require background knowledge before modelling. Consider a snippet from the CLDW related to one of the landmarks named Pooley Bridge with NER tagging represented in figure 1 (The NER and semantic tagging system for extracting spatial entities has been developed by our colleagues at Lancaster as part of the first module [4]). Here, an interpretation of possible relationships and ambiguous terms is presented for a few sentences:

► **Example 1.** *From Penrith two roads lead to Pooley Bridge, about six miles distant, which spans the Eamont just at its issue from Ulleswater. Either road may be taken, be we recommend that which follows the Shap road to Eamont Bridge. Carleton Hall is near to it on the left. Cross the bridge, and take the first road to the right. At this point, on the left, are the druidical remains called King Arthur's Round Table and Mayborough. Immediately after crossing Pooley Bridge, the road runs along the western shore of Ulleswater to Patterdale, a distance of ten miles; but, before proceeding along it, the tourist would do well to take a walk of a few miles along the eastern shore, in the direction of Martindale, from several points on which he will obtain a good view of the lake.*



■ **Figure 1** NER and semantic tagging system developed by our project colleagues I.Ezeani and P.Rayson [4].

Sentence 1: From Penrith two roads lead to Pooley Bridge, about six miles distant, which spans the Eamont just at its issue from Ulleswater.

QSR-based relations and interpretation:

- **place(penrith)**
- **place(pb)** : Pooley Bridge, pb, is both a bridge and a town, one needs to consider both, and have two different logical names
- **distance(penrith, pb, about(6), miles)** : approximate measurements with about(), a reasoning mechanism could be developed for “aboutness”
- **road(road1)** : “roads” are really “routes” since there are various roads and road segment which constitute them
- **road(road2)** : one could explicitly say that road1 and road2 are 6 miles long or possibly have a rule which says that if r is a road from x to y and the distance between x and y is z, then r must be at least z long

- **end(road1,penrith)** : a separate start(,) relation would be required if road is to be made oriented
- **end(road1,pb)**
- **end(road2,penrith)**
- **end(road2,pb)** : a further statement could be added to note the implicit fact that there are *only* 2 roads from penrith to pb
- **bridge(pb)**
- **spans(pb,eamont)** : some rules could be added about spanning and bridges that lets one infer that one can get from one side to the other via the bridge, and also one can only cross a river via a bridge or a tunnel or a ford. Moreover, all bridges span something and have two ends
- **river(eamont)** : not explicit in the text, requires background knowledge
- **source(eamont,ullswater)**
- **lake(ullswater)** : not explicit in the text, requires background knowledge

Sentence 2: Carleton Hall is near to it on the left.

QSR-based relations and interpretation:

- **near(carleton Hall,pb)** : “near” is a vague term and one has to separately consider what axioms/rules might apply to it in this kind of geographical context. Vague spatial terms have received a lot of attention in the literature but without any definitive treatment. Note that “near” is not transitive, i.e. from near(a,b) and near(b,c) we cannot conclude near(a,c).
- **direction(carleton Hall, left eamont bridge,shap)** : Directions provide different frames of references. Here, a direction predicate is defined with four arguments: the “figure” (i.e. the thing being pointed out), the direction (here left), the “ground” (i.e. the place from where the direction is being pointed out from), and the direction the person pointing is facing. Note that if roads have a start and an end then fourth argument is not required. Almost certainly a number of different direction predicates for the different situations are needed.

4 Conclusion

This paper presents a small part of a larger research project which aims to uncover spatial and temporal dynamics of narratives. The research will create a step-change in the way we explore the geographies described in large textual collections by exploring how GIS and related tools can identify, analyse and visualise qualitative senses of place alongside the quantitative spatial information more typically used in geographical information science (GISc). This will allow us to redefine the way that computer technologies represent place and transform the abilities of social scientists and humanists to understand and interpret narrative accounts about place.

References

- 1 John Agnew and David N. Livingstone. *The SAGE Handbook of Geographical Knowledge*. SAGE Publications Ltd, London, 2011. doi:10.4135/9781446201091.
- 2 Eliseo Clementini. A conceptual framework for modelling spatial relations. *Information Technology and Control*, 48:5–17, 2019. doi:10.5755/j01.itc.48.1.22246.
- 3 Christopher Elliott Donaldson, Ian N. Gregory, and Joanna E. Taylor. Implementing corpus analysis and gis to examine historical accounts of the english lake district. in: Historical atlas. In Peter Bol, editor, *Historical atlas: its concepts and methodologies*, pages 152–172. Northeast

- Asian History Foundation, Seoul, 2017. URL: <https://eprints.lancs.ac.uk/id/eprint/81425/>.
- 4 Ignatius Ezeani, Paul Rayson, and Ian N. Gregory. Extracting imprecise geographical and temporal references from journey narratives. In *Proceedings of Text2Story — Sixth Workshop on Narrative Extraction From Texts*, 2023. URL: <https://ceur-ws.org/>.
 - 5 Erum Haris, Keng Hoon Gan, and Tien-Ping Tan. Spatial information extraction from travel narratives: Analysing the notion of co-occurrence indicating closeness of tourist places. *Journal of Information Science*, 46(5):581–599, 2020. doi:10.1177/0165551519837188.
 - 6 Gregor Jossež, Klaus Arthur Schmid, Andreas Züfle, Georgios Skoumas, Matthias Schubert, Matthias Renz, Dieter Pfoser, and Mario A. Nascimento. Knowledge extraction from crowd-sourced data for the enrichment of road networks. *Geoinformatica*, 21(4):763–795, 2017. doi:10.1007/s10707-017-0306-1.
 - 7 Parisa Kordjamshidi, Martijn Van Otterlo, and Marie-Francine Moens. Spatial role labeling: Towards extraction of spatial relations from natural language. *ACM Trans. Speech Lang. Process.*, 8(3), 2011. doi:10.1145/2050104.2050105.
 - 8 Parisa Kordjamshidi, Martijn Otterlo, and Marie-Francine Moens. From language towards formal spatial calculi. In *Proceedings of the Workshop on Computational Models of Spatial Language Interpretation at Spatial Cognition*, volume 620, pages 17–24, 2010. URL: <http://ceur-ws.org/Vol-620>.
 - 9 Molly Miranker and Alberto Giordano. Text mining and semantic triples: Spatial analyses of text in applied humanitarian forensic research. *Digital Geography and Society*, 1:100005, 2020. doi:10.1016/j.diggeo.2020.100005.
 - 10 Patricia Murrieta-Flores, Mariana Favila-Vázquez, and Aban Flores-Morán. Spatial humanities 3.0: Qualitative spatial representation and semantic triples as new means of exploration of complex indigenous spatial representations in sixteenth century early colonial mexican maps. *International Journal of Humanities and Arts Computing*, 13(1-2):53–68, 2019. doi:10.3366/ijhac.2019.0231.
 - 11 Space Time Narratives Project. Understanding imprecise space and time in narratives through qualitative representations, reasoning, and visualisation. <https://spacetime narratives.github.io/>, 2023.
 - 12 Robert Smail, Ian N.Gregory, and Joanna E.Taylor. Qualitative geographies in digital texts: Representing historical spatial identities in the lake district. *International Journal of Humanities and Arts Computing*, 13(1-2):28–38, 2019. doi:10.3366/ijhac.2019.0229.
 - 13 Joanna E. Taylor and Ian N. Gregory. *Deep Mapping the Literary Lake District*. Bucknell University Press, 2022. doi:10.2307/j.ctv2v55bsf.

Exascale Agent-Based Modelling for Policy Evaluation in Real-Time (ExAMPLER)

Alison Heppenstall¹  

School of Social and Political Sciences, University of Glasgow, UK

J. Gary Polhill  

The James Hutton Institute, Aberdeen, UK

Mike Batty  

Bartlett Centre for Advanced Spatial Analysis, University College London, UK

Matt Hare  

The James Hutton Institute, Aberdeen, UK

Doug Salt  

The James Hutton Institute, Aberdeen, UK

Richard Milton  

Bartlett Centre for Advanced Spatial Analysis, University College London, UK

Abstract

Exascale computing can potentially revolutionise the way in which we design and build agent-based models (ABM) through, for example, enabling scaling up, as well as robust calibration and validation. At present, there is no exascale computing operating with ABM (that we are aware of), but pockets of work using High Performance Computing (HPC). While exascale computing is expected to become more widely available towards the latter half of this decade, the ABM community is largely unaware of the requirements for exascale computing for agent-based modelling to support policy evaluation. This project will engage with the ABM community to understand what computing resources are currently used, what we need (both in terms of hardware and software) and to set out a roadmap by which to make it happen.

2012 ACM Subject Classification Computing methodologies → Modeling and simulation

Keywords and phrases Exascale computing, Agent-Based Modelling, Policy evaluation

Digital Object Identifier 10.4230/LIPIcs.GIScience.2023.38

Category Short Paper

Funding Supported by the Engineering and Physical Sciences Research Council (project reference EP/Y008839/1) and the Scottish Government Rural and Environment Science and Analytical Services Division (project reference JHI-C5-1). AH was supported by grants from UKPRP (MR/S037578/2), Medical Research Council (MC_UU_00022/5) and Scottish Government Chief Scientist Office (SPHSU20).

1 The transformative potential of exascale computing for agent-based modelling

Exascale computing is defined as computing capable of 10^{18} floating-point operations per second (FLOPS). Though instructions executed per second is of greater relevance to agent-based modelling than purely floating-point operations, the two are approximately the same. A recent experiment with an agent-based model conducted by the authors used 76 CPU

¹ Corresponding author



days of computing time on a cluster with mean 4390 “bogomips” (“bogos” (i.e. approximate) million instructions per second) CPUs. This is approximately 3×10^{16} CPU instructions, which an exascale computer could theoretically complete in three hundredths of a second, in comparison with roughly a working day (8 hours or so) on a 200 CPU high-performance computing cluster. This is a six orders of magnitude improvement in computing time.

The potential benefits to agent-based modellers of access to exascale computing are immediate, even based on existing practice. These would be manifested most trivially in being able to sample models’ high-dimensional parameter spaces more densely during calibration. Since exascale computers are massively parallel architectures, there is also immediate potential in appropriately parallelised larger-scale simulations enabling us to model megacities and countries with millions or billions of agents; though here there are challenges in taking full advantage of exascale computing because of thread-blocking and shared memory issues.

However, this huge gain in computing power has rather more revolutionary potential for agent-based modelling than merely doing what we already do, but bigger. There are three main activities occupying a significant time in empirical agent-based modelling. First is assembling and preparing data; second is designing and building the agent-based model itself and any supporting software; third is running the simulation experiments for calibration, validation and scenario analysis and processing and visualizing the outputs. The third of these – assuming the tools already exist – is most trivially addressed by exascale computing: a process that takes days can be completed in a few seconds. The first two, which are less embarrassing in duration when experimentation takes so long, then start to look rather more embarrassing.

Squazzoni et al.’s [6] call for improvements in data sharing and modelling practice in the early stages of the COVID crisis are no less relevant now than they were then. Though COVID brought some of the benefits of agent-based modelling into sharp focus as authors such as Thompson et al. [7] and Badham et al. [3] worked with policymakers to evaluate scenarios for managing the crisis. With an *existing* agent-based model, data, and analysis and visualisation tools, exascale computing could already support a creative, transdisciplinary discussion about how to handle a developing emergency. Such a discussion would be greatly enhanced, however, if the model could be adapted and new data brought in, as people became aware of potential cascading consequences of their interventions. With appropriate software and institutional support, enhancements like these could be realised, significantly improving the attractiveness of bringing agent-based modelling in to such conversations.

2 Challenges

Existing work with high-performance computing (HPC) infrastructure in agent-based modelling makes it clear that realising the potential of exascale computing in the area will not be without its challenges. These all largely pertain to accessibility. There are three main areas to consider: technical, institutional, and cultural. Much of the following is anecdotal, but the points will be familiar to those who have tried to access HPC to run their agent-based models.

From a technical perspective, Alessa et al. [1] put out their “All Hands” call to create a community of practice around social simulation and cyberinfrastructure in 2006, referring explicitly to the fact that developing agent-based models, even on platforms such as NetLogo, entails a learning curve that is a significant barrier to adoption of agent-based modelling in the social sciences. Introducing a special issue of *JASSS* on “grand challenges” more than

Please see notes in the Service Specification document regarding the maximum amounts of time that can be applied for and technical specifications.

	Largest Job	Typical Job	Smallest Job
Number of nodes	(Please Complete Table)		
Number of cores/GPUs used per node			
Wallclock time for each job*			
Number of jobs of this type			
Memory per node required.			

*The maximum permitted wallclock time per job is a function of local Service centre policy.

■ **Figure 1** A screenshot from one of the EPSRC’s “Technical Assessment” forms to access national computing infrastructure.

ten years later, An et al. [2] are still in a position to refer to the “steep learning curves” (para. 3.2 – note the plural) faced by modelling novices. Accessing HPC currently involves command-line interfaces, shell scripts, and SSH (Secure Shell Protocol) arcanery that social scientists are not desperately keen to learn. This does not mean that they do not want to use the technology. Their primary interest in doing so, however, is in the practical benefits to them in the insights gained for their case study. Social scientists don’t necessarily get their intellectual “kicks” from playing with advanced technology. This means HPC needs to be easy to use – ideally (though impossibly) to such an extent that there is a button on the interface of their modelling tool that says “Run this experiment on HPC”.

Institutionally, accessing HPC infrastructure is surrounded by gatekeepers, who need forms to be filled, stipulating information that social scientists may not be in a position to provide. For example, in the UK, national high-end computing infrastructure access is managed by the Engineering and Physical Sciences Research Council. This is managed by calls with deadlines², which require applicants to complete “Technical Assessment” forms stipulating information such as that in figure 1. From the perspective of managing the HPC, this kind of information makes sense in that it helps with planning usage of the machines to ensure that everyone’s needs can be met. Further, individuals running the facilities can be extremely helpful to the first-time user in advising on how to complete these forms. However, Polhill [5] has pointed out that agent-based modellers may not be able to provide accurate estimates of run-time or memory demand from running the models, for sound theoretical computer science reasons *that anyone with a degree in computer science should know*.

The cultural side is somewhat harder to articulate, and could be ironically phrased as the question of whether your code is “worthy” of the very expensive computing equipment on which you hope to run it. The social sciences suffer perennially from physics envy, but physicists regularly use HPC as part of work on particle collision and cosmology, some even claiming to be in search of the “mind of God” [4, p. 175]. How can the social sciences compete with such majesty? Rather more prosaically, computer scientists are not especially excited about “embarrassingly parallel” problems such as running models repeatedly to explore parameter space. Massive distributed models with difficult multi-threading memory and

² e.g. <https://www.ukri.org/councils/epsrc/facilities-and-resources/using-epsrc-facilities-and-resources/apply-for-access-to-high-performance-computing-facilities/>

CPU co-ordination problems that don't break the benefits of parallelism are more interesting to them. This is strange because they have built machines that are brilliant and hugely efficient for the former kind of problem, but aren't really designed to do the latter nearly so well. Perhaps most problematically (and mundanely), however understandable it may be, the worthiness of your code is also reflected in its efficiency – have you used the right data structures and algorithms to reach the results of running the code with as few instructions executed as possible? Some might simply be pleased that they've got a model that runs without crashing...

3 Benefits beyond exascale

In thinking about the software, institutional and data support that empirical agent-based modellers need to take full advantage of the potential of exascale computing, there are opportunities to think about wider benefits to the community. By being involved in the conversation about HPC access, and making the case for our requirements, we may be able to break down some of the barriers described above, reaching a point whereby HPC use is more routine in agent-based modelling work. The software developed to support exascale agent-based modelling could also be useful for agent-based modelling on a laptop – especially if we are somehow able provide tools that enable agent-based models to be built rapidly.³ If there are ways of easing access to data suitable for using in empirical agent-based models, and learning from and building on others' experiences with doing so, then this will advance empirical applications of agent-based models by reducing the time investment this modelling step currently requires.

References

- 1 Lilian Na Alessa, Melinda Laituri, and Michael Barton. An “all hands” call to the social science community: Establishing a community framework for complexity modeling using agent based models and cyberinfrastructure. *Journal of Artificial Societies and Social Simulation*, 9(4):6, 2006. URL: <https://www.jasss.org/9/4/6.html>.
- 2 Li An, Volker Grimm, and Billie L. Turner II. Editorial: Meeting grand challenges in agent-based models. *Journal of Artificial Societies and Social Simulation*, 23(1):13, 2020. doi:10.18564/jasss.4012.
- 3 Jennifer Badham, Pete Barbrook-Johnson, Camila Caiado, and Brian Castellani. Justified stories with agent-based modelling for local COVID-19 planning. *Journal of Artificial Societies and Social Simulation*, 24(1):8, 2021. doi:10.18564/jasss.4532.
- 4 Stephen W. Hawking. *A Brief History of Time: From the Big Bang to Black Holes*. Bantam Press, London, UK, 1988.
- 5 Gary Polhill. Antisocial simulation: using shared high-performance computing clusters to run agent-based models. *Review of Artificial Societies and Social Simulation*, 14 December 2022, 2022. URL: <https://rofasss.org/2022/12/14/antisoc-sim/>.
- 6 Flaminio Squazzoni, Gary Polhill, Bruce Edmonds, Petra Ahrweiler, Patrycja Antosz, Geeske Scholz, Émile Chappin, Melania Borit, Harko Verhagen, Francesca Giardini, and Nigel Gilbert. Computational models that matter during a global pandemic outbreak: A call to action. *Journal of Artificial Societies and Social Simulation*, 23(2):10, 2020. doi:10.18564/jasss.4298.

³ The “reusable building blocks” work, such as that being run by CoMSES.Net is a case in point. See <https://github.com/comses-model-building-blocks>

- 7 Jason Thompson, Rod McClure, Tony Blakely, Nick Wilson, Michael G. Baker, Jasper S. Wijnands, Thiago Herick De Sa, Kerry Nice, Camilo Cruz, and Mark Stevenson. Modelling SARS-CoV-2 disease progression in Australia and New Zealand: an account of an agent-based approach to support public health decision-making. *Australian and New Zealand Journal of Public Health*, 46(3):292–303, 2022. doi:10.1111/1753-6405.13221.

A Hierarchical and Geographically Weighted Regression Model and Its Backfitting Maximum Likelihood Estimator

Yigong Hu¹  

School of Geographical Sciences, University of Bristol, UK

Richard Harris  

School of Geographical Sciences, University of Bristol, UK

Richard Timmerman  

School of Geographical Sciences, University of Bristol, UK

Binbin Lu  

School of Remote Sensing and Information Engineering, Wuhan University, Hubei, China

Abstract

Spatial heterogeneity is a typical and common form of spatial effect. Geographically weighted regression (GWR) and its extensions are important local modeling techniques for exploring spatial heterogeneity. However, when dealing with spatial data sampled at a micro-level but the geographical locations of them are only known at a higher level, GWR-based models encounter several problems, such as difficulty in establishing the bandwidth. Because data with this characteristic exhibit spatial hierarchical structures, such data can be suitably handled using hierarchical linear modeling (HLM). This model calibrates random effects for sample-level variables in each group to address spatial heterogeneity. However, it does not work when exploring spatial heterogeneity in some group-level variables when there is insufficient variance in each group. In this study, we therefore propose a hierarchical and geographically weighted regression (HGWR) model, together with a back-fitting maximum likelihood estimator, that can be applied to examine spatial heterogeneity in the regression relationships of data where observations nest into high-order groupings and share the same or very close coordinates within those groups. The HGWR model divides coefficients into three types: local fixed effects, global fixed effects, and random effects. Results of a simulation experiment show that HGWR distinguishes local fixed effects from others and also global effects from random effects. Spatial heterogeneity is reflected in the estimates of local fixed effects, along with the spatial hierarchical structure. Compared with GWR and HLM, HGWR produces estimates with the lowest deviations of coefficient estimates. Thus, the ability of HGWR to tackle both spatial and group-level heterogeneity simultaneously suggests its potential as a promising data modeling tool for handling the increasingly common occurrence where data, in secure settings for example, remove the specific geographic identifiers of individuals and release their locations only at a group level.

2012 ACM Subject Classification Information systems → Geographic information systems

Keywords and phrases spatial modelling, hierarchical data, spatial heterogeneity, geographically weighted regression

Digital Object Identifier 10.4230/LIPICs.GIScience.2023.39

Category Short Paper

Supplementary Material *Software (Source code)*: <https://github.com/hpdell/hgwr>
archived at `swh:1:dir:c9c2bab2a6428b8d3b6d25a3da472653018a7fae`

Text (Blog post): <https://hpdell.github.io/GIScience-Materials/posts/HGWR/>

Funding *Yigong Hu*: Yigong Hu was sponsored by the China Scholarship Council with the University of Bristol (No. 202106270029).

¹ Corresponding author.



© Yigong Hu, Richard Harris, Richard Timmerman, and Binbin Lu;
licensed under Creative Commons License CC-BY 4.0

12th International Conference on Geographic Information Science (GIScience 2023).

Editors: Roger Beecham, Jed A. Long, Dianna Smith, Qunshan Zhao, and Sarah Wise; Article No. 39; pp. 39:1–39:6
Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Introduction

In statistics and data analysis, regression models are powerful tools in examining relationships in data. However, the ordinary linear regression, as a model of global relationships, holds many limitations in dealing with spatial data [5] because the relationship between variables may not keep constant across the whole area. In spatial statistics, this phenomenon is called “spatial heterogeneity” [2]. To uncover such an effect, many local-form spatial statistic methods are proposed to discover underlying spatial heterogeneity in data [5]. The geographically weighted regression (GWR) [3] model and its extensions are popular ones. These methods calibrate a unique model at each location to produce spatially varying coefficients by borrowing samples from its geographical neighbors defined by spatial distances. Shorter distance gives rise to higher weighting. Among its extensions, the multiscale GWR (MGWR) [6, 9] has many attractive features. MGWR specifies a unique bandwidth for each coefficient to improve the goodness of fit and prediction accuracy [9]. The hierarchical linear model [10], is also an important method for finding spatial heterogeneity in data of hierarchical structure. When samples are grouped by their locations, HLM calibrates some effects for samples in each group (called “random effects”) to fit for spatially varying relationships, whereas other effects are treated as “fixed effects” that are constant for all groups [8].

In recent years, spatially hierarchical data have become increasingly popular in real world analysis since samples can be naturally nested in different spatial scales. For example, in the Biobank database [1] which consists of health information from 0.5 million UK participants, their addresses are nested into 1km-by-1km grid cells to protect their privacy. With the development of spatial big data and improved access to administrative data through secure data settings, it is increasingly common to find data sets where the attributes of the sample are available at a different geographic scale to their geographical identifiers. In spatial data of hierarchical structures, effects of variables may work in different ways. For example, group-level variables – that keep constant within groups – may have global or local effects, and sample-level variables are the same. No matter which variables, the basic GWR model always treat their effects as local ones, and estimate them by data borrowing from geographical neighbors. When dealing with group-level variables, the repeated values increase the risk of singular matrix. MGWR works similarly, only that it assigns variable-specified bandwidth settings and variables of global effects will be assigned a huge bandwidth up to infinity to estimate global effects. Fixed effects and random effects in HLM can be used to discover global and local effects, respectively. Fixed effects can be estimated for both group-level variables and sample-level variables. However, random effects only work for sample-level variables, which vary among individuals as opposed to the group-level ones. Because values of group-level variables are determined by their locations. Thus, there is no sufficient variation to calibrate random effects for them within each group. We need a special method to properly estimate effects of the variables with spatial heterogeneity.

In this article, we propose a hierarchical and geographically weighted regression (HGWR) model and its estimator based on backfitting and maximum likelihood (BFML) algorithms to solve the above-mentioned issues. This model calibrates two types of fixed effects – local fixed effects and global fixed effects – and random effects. We conducted a simulation experiment to ascertain whether HGWR could successfully distinguish local effects from other effects. We also compared its performance with GWR, MGWR, and HLM.

2 Model

The HGWR model is designed for data with a spatially hierarchical structure. In a data set with n samples divided in m groups according to their locations, the variance of dependent variable \mathbf{y} can be explained with the following three parts: local-fixed effects $\boldsymbol{\gamma}$ for variables

\mathbf{G} that vary with location; global-fixed effects β for variables X that are constant across the whole area; and random effects μ for variables Z that vary from group to group. The model for sample j in group i can be expressed as Equation 1,

$$y_{ij} = \mathbf{G}_i\gamma_i + \mathbf{X}_i\beta + \mathbf{Z}_{ij}\mu_i + \epsilon_{ij} \tag{1}$$

where γ_i , β_i and μ_i are coefficients of local fixed effects, global fixed effects, and random effects respectively; ϵ_{ij} is the remaining random error. Then this model can be written in a matrix form as Equation 2,

$$\mathbf{y} = \text{diag}\{\mathbf{G}\boldsymbol{\gamma}\} + \mathbf{X}\beta + \mathbf{Z}\boldsymbol{\mu} + \boldsymbol{\epsilon} \tag{2}$$

where $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \dots, \gamma_m)$,

$$\mathbf{G} = \begin{pmatrix} \mathbf{G}_1 \\ \mathbf{G}_2 \\ \vdots \\ \mathbf{G}_m \end{pmatrix}, \mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_m \end{pmatrix}, \mathbf{Z} = \begin{pmatrix} \mathbf{Z}_1 & & & \\ & \mathbf{Z}_2 & & \\ & & \ddots & \\ & & & \mathbf{Z}_m \end{pmatrix}, \boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_m \end{pmatrix},$$

$\boldsymbol{\mu} \sim N(0, \sigma^2 \mathbf{D})$, and $\boldsymbol{\epsilon} \sim N(0, \sigma^2 \mathbf{I})$, and $\mathbf{G}\boldsymbol{\gamma}$ here is regarded as a product of block matrices, such that

$$\mathbf{G}\boldsymbol{\gamma} = \begin{pmatrix} \mathbf{G}_1\gamma_1 & \mathbf{G}_1\gamma_2 & \cdots & \mathbf{G}_1\gamma_m \\ \mathbf{G}_2\gamma_1 & \mathbf{G}_2\gamma_2 & \cdots & \mathbf{G}_2\gamma_m \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{G}_m\gamma_1 & \mathbf{G}_m\gamma_2 & \cdots & \mathbf{G}_m\gamma_m \end{pmatrix}, \text{diag}\{\mathbf{G}\boldsymbol{\gamma}\} = \begin{pmatrix} \mathbf{G}_1\gamma_1 \\ \mathbf{G}_2\gamma_2 \\ \vdots \\ \mathbf{G}_m\gamma_m \end{pmatrix}.$$

In this model, coefficients $\gamma_1, \gamma_2, \dots, \gamma_m$ are estimated group-by-group as for other GWR models using weighted least squared estimation [3] with a uniform bandwidth.

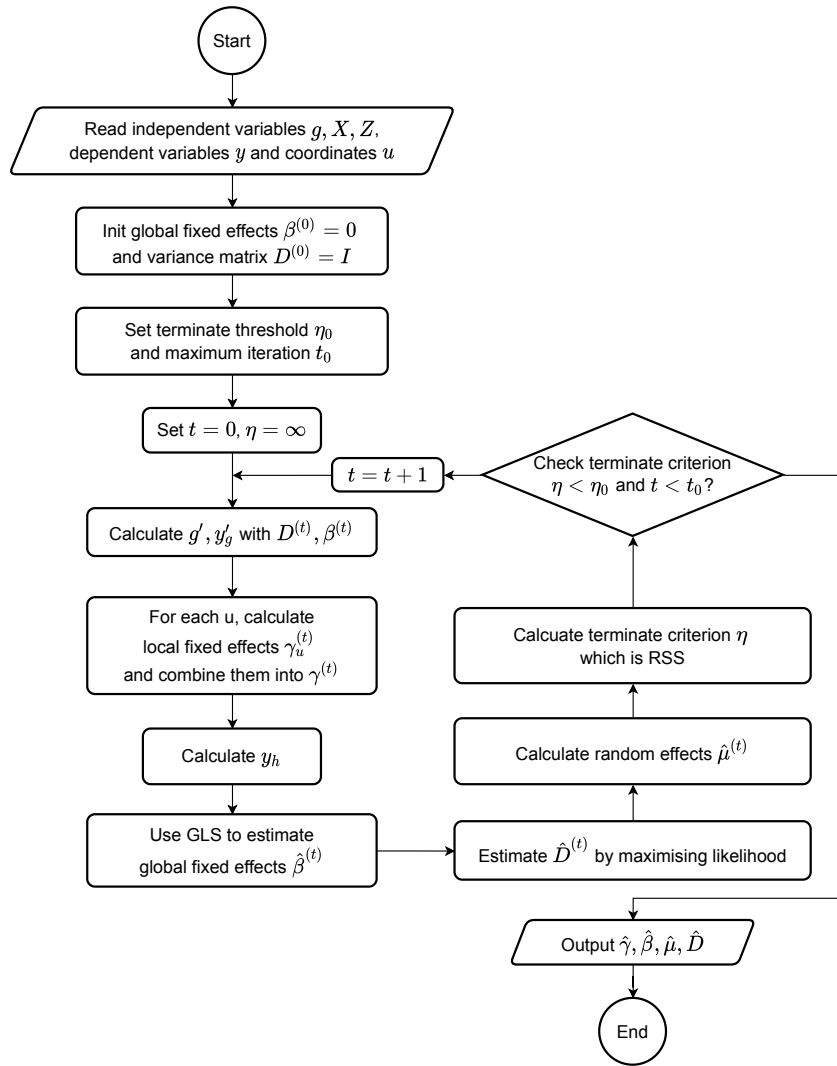
A back-fitting procedure, shown in Figure 1, can be applied to estimate parameters in this model following a similar methodical approach to [4]. In this workflow, when calibrating local fixed effects $\hat{\gamma}^{(t)}$ in each iteration, the algorithm can optimize the bandwidth value via golden-selection [7] according to the CV criterion. This algorithm is very efficient and effective in minimizing univariate functions.

3 Simulation Experiments

To evaluate the performance of HGWR and compare this model with HLM, GWR and MGWR, some simulation experiments ² are designed. In particular, the performance was measured regarding the ability to properly distinguish local fixed effects from global fixed effects under the circumstance that random effects exist.

A spatial data set of 21,434 random samples was generated that were unevenly spread across 625 locations. The data generating process was inspired by [6]. For each data point, four independent variables ($\mathbf{g}_1, \mathbf{g}_2, \mathbf{x}_1, \mathbf{z}_1$) were generated according to the standard multivariate normal distribution. To simulate group-level spatial-related variables, the mean of \mathbf{g}_1 and \mathbf{g}_2 at each location were substituted for the original values. Samples located together share coefficient values. Values of the generated coefficients are shown in the first row of Figure 2. Results of the for models are shown in other rows.

² Please turn to <https://hpdell.github.io/GIScience-Materials/posts/HGWR/> for more details.

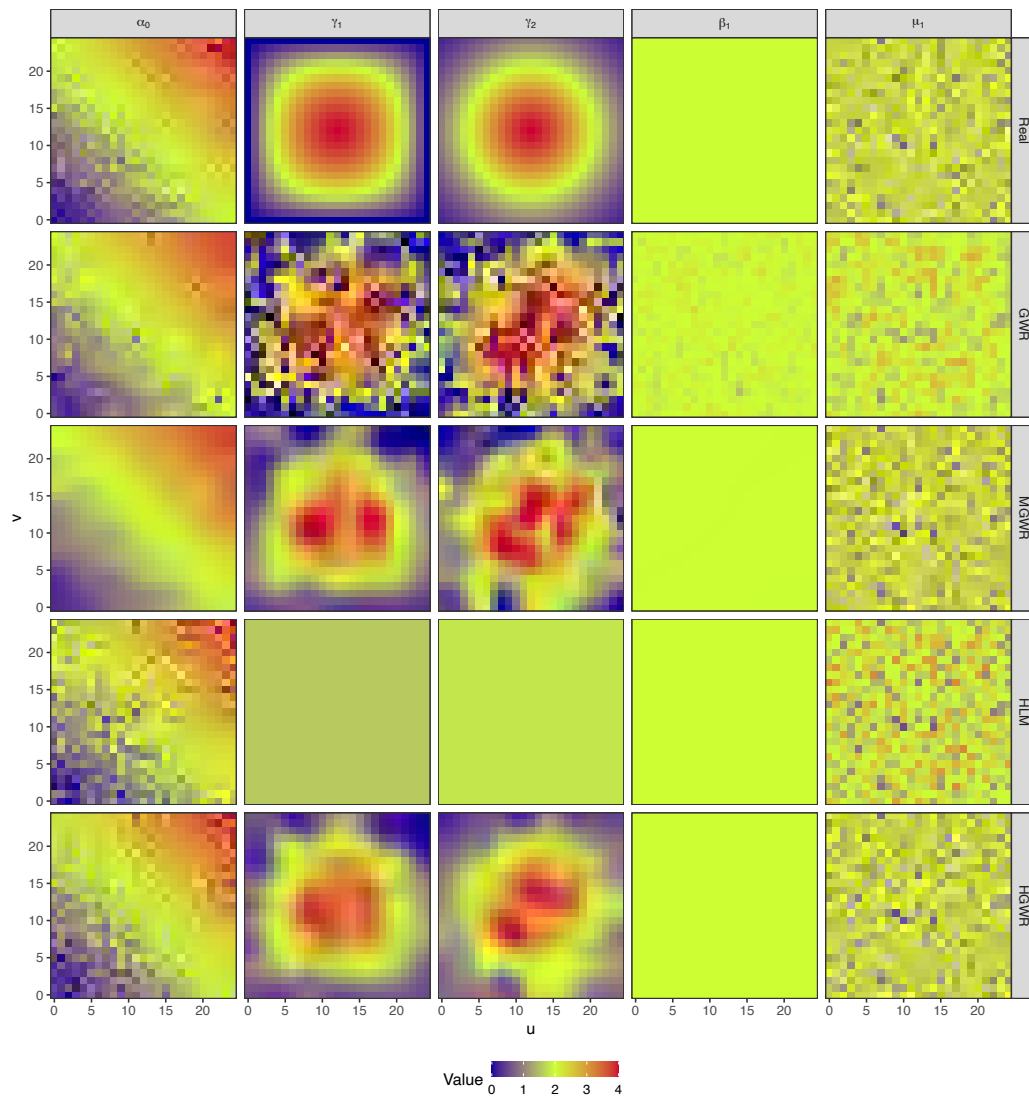


■ **Figure 1** Diagram of the BFML estimator for HGWR, where $RSS = (\mathbf{y} - \hat{\mathbf{y}})^T(\mathbf{y} - \hat{\mathbf{y}})$ and $\hat{\mathbf{y}} = \mathbf{G}\hat{\boldsymbol{\gamma}} + \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{Z}\hat{\boldsymbol{\mu}}$.

In the results of GWR, spatial heterogeneity is revealed in estimates for all variables. Although $\hat{\beta}_1$ should be constant across the study area, GWR still generate spatially varying estimates for it. This is a kind of over-fitting from the spatial perspective. However, for estimates of μ_1 , they are smoothed compared with actual values, even though the bandwidth selected is small enough. Because the bandwidth is small, estimates for γ_1 and γ_2 are too local. Consequently, there are quite a few outliers disrupting the spatial trend.

MGWR partly gets over issues of GWR by adopting parameter-specified bandwidths, instead of a uniform bandwidth. It performs better when estimating γ_1 and γ_2 . For global fixed effects, MGWR still generates spatially varying estimates, but they vary more slightly than estimates from GWR. For random effects, the results are slightly smoothed as well. Besides, MGWR it requires a lot of computing time and memory.

In the results of HLM, there is only one estimate for β_1 across the whole area as well as estimates for μ_1 , the problem lies in estimates for γ_1 and γ_2 . As they are fixed effects in HLM, their estimates are also constant for all samples. However, spatial heterogeneity is expected in them.



■ **Figure 2** Real values and estimated values.

HGWR is the final solution. For global fixed effects, it generates globally constant estimates for all samples. For random effects, it does not smooth the estimates because they are not obtained by borrowing points. And for local fixed effects, we can discover spatial heterogeneity from their estimates. And it does not repeat computation for samples at each location. Computationally, it is more efficient because it does not repeat geographically weighted fitting at every sample within a higher-level group where models are the same. On the dataset used in the experiment, calibrating the HGWR model only took 6.06 seconds, which reduced the calculation time by 4 minutes compared to GWR (3.55 mins); and reduced it by nearly 4.4 hours compared to MGWR (4.41 hours) paralleled by 48 threads. These findings have been double-checked via repeating the experiment 100 times.

4 Conclusion

In this article we proposed a BFML estimator for a HGWR model. Compared with HLM, this method divides fixed effects into global and local effects. For local fixed effects, this model applies a spatial heterogeneity assumption and estimates the effects using the GWR method. For global fixed effects and random effects, this model adopts a similar method as in HLM, i.e., maximum likelihood. To facilitate cooperation between the two methods, a back-fitting procedure was developed. It is demonstrated that HGWR can properly estimate local fixed effects, global fixed effects, and random effects simultaneously. HGWR can successfully distinguish local fixed effects from other effect types. For local fixed effects, spatial heterogeneity is considered as with GWR; moreover, global fixed effects and random effects are estimated as accurately as when using HLM. Thus, HGWR can be regarded as a successful combination of GWR and HLM. In this stage, there are some limitations remaining to be solved, such as convergence conditions and statistical inferences. Nevertheless, with the popularity of spatiotemporal big data, situations wherein the specific parameters for which HGWR was optimized are becoming more prevalent, suggesting that HGWR holds considerable promise as a useful tool for analyzing such data sets.

References

- 1 Naomi E. Allen, Cathie Sudlow, Tim Peakman, and Rory Collins. Uk biobank data: Come and get it. *Science Translational Medicine*, 6(224):1–4, February 2014. doi:10.1126/scitranslmed.3008601.
- 2 Luc Anselin. *Spatial Econometrics: Methods and Models*. Number 4 in Studies in operational regional science. Kluwer Academic Publishers, Dordrecht, Boston, 1988.
- 3 Chris Brunsdon, A. Stewart Fotheringham, and Martin E. Charlton. Geographically weighted regression: A method for exploring spatial nonstationarity. *Geographical Analysis*, 28(4):281–298, February 1996. doi:10.1111/j.1538-4632.1996.tb00936.x.
- 4 Feng Chen, Yee Leung, Chang-Lin Mei, and Tung Fung. Backfitting estimation for geographically weighted regression models with spatial autocorrelation in the response. *Geographical Analysis*, 54(2):357–381, April 2022. doi:10.1111/gean.12289.
- 5 A. Stewart Fotheringham and Chris Brunsdon. Local forms of spatial analysis. *Geographical Analysis*, 31(4):340–358, September 2010. doi:10.1111/j.1538-4632.1999.tb00989.x.
- 6 A. Stewart Fotheringham, Wenbai Yang, and Wei Kang. Multiscale geographically weighted regression (mgwr). *Annals of the American Association of Geographers*, 107(6):1247–1265, November 2017. doi:10.1080/24694452.2017.1352480.
- 7 Dorothy Margaret Greig. *Optimisation*. Longman Publishing Group, 1980.
- 8 A. M. Latimer, S. Banerjee, H. Sang Jr, E. S. Mosher, and J. A. Silander Jr. Hierarchical models facilitate spatial analysis of large data sets: a case study on invasive plant species in the northeastern united states. *Ecology Letters*, 12(2):144–154, February 2009. doi:10.1111/j.1461-0248.2008.01270.x.
- 9 Binbin Lu, Chris Brunsdon, Martin Charlton, and Paul Harris. Geographically weighted regression with parameter-specific distance metrics. *International Journal of Geographical Information Science*, 31(5):982–998, May 2017. doi:10.1080/13658816.2016.1263731.
- 10 Stephen W. Raudenbush. Hierarchical linear models and experimental design. In *Applied Analysis of Variance in Behavioral Science*, pages 459–496. L. K. Edwards, 1993.

Introducing a General Framework for Locally Weighted Spatial Modelling Based on Density Regression

Yigong Hu¹  

School of Geographical Sciences, University of Bristol, UK

Binbin Lu  

School of Remote Sensing and Information Engineering, Wuhan University, Hubei, China

Richard Harris  

School of Geographical Sciences, University of Bristol, UK

Richard Timmerman  

School of Geographical Sciences, University of Bristol, UK

Abstract

Traditional geographically weighted regression and its extensions are important methods in the analysis of spatial heterogeneity. However, they are based on distance metrics and kernel functions compressing differences in multidimensional coordinates into one-dimensional values, which rarely consider anisotropy and employ inconsistent definitions of distance in spatio-temporal data or spatial line data (for example). This article proposes a general framework for locally weighted spatial modelling to overcome the drawbacks of existing models using geographically weighted schemes. Underpinning it is a multi-dimensional weighting scheme based on density regression that can be applied to data in any space and is not limited to geographic distance.

2012 ACM Subject Classification Information systems → Geographic information systems

Keywords and phrases Spatial heterogeneity, Multidimensional space, Density regression, Spatial statistics

Digital Object Identifier 10.4230/LIPIcs.GIScience.2023.40

Category Short Paper

Supplementary Material *Software (Source code)*: <https://github.com/GWmodel-Lab/GWmodel13>
archived at `swh:1:dir:24841fa8fac1919085decceb53131f35634b6b01`

Funding *Yigong Hu*: Yigong Hu was sponsored by the China Scholarship Council with the University of Bristol (No. 202106270029).

1 Introduction

In recent years, analysis of spatial heterogeneity – for example, spatially varying regression relationships – has attracted increasing interest from researchers. Among the local-form spatial modelling methods, geographically weighted regression (GWR) [1] is popular. It fits a unique weighted least squared model at multiple locations across a study region by borrowing points from each location’s geographic neighbours. Extensions include geographically and temporally weighted regression (GTWR) [2], enhancing basic GWR’s ability to model more kinds of data. Basic GWR, on 2D spatial data sets, uses weights based on geographic distances between samples. Extended versions may adapt the weights to incorporate other

¹ Corresponding author.



kinds of “distance” but are still rooted back into one-dimensional distance metrics. This raises the problem of how to compress differences in multidimensional coordinates into a one-dimensional distance value.

Additionally, even when the metric is simple, differences in geographic scales of different dimensions may cause unexpected problems. This phenomenon is called “anisotropy”. For example, the range of vertical distances is generally different from that of horizontal distances. Consequently, when we incorporate distance in the 3D space to weight samples, relatively large changes in heights may present very limited effects on weights (without rescaling the vertical distances, at least). The problem is more evident when time is considered as this is, of course, measured in units of time, not of space. They are not directly compatible. These problems highlight the limitation of reducing multidimensional spaces into a single-dimensional weighting based on some notion of “closeness” or least distance.

In this paper, we introduce a general framework for locally weighted geographic and other spatial modelling based on density regression (DLSM: density-based local spatial models). This model essentially follows the workflow of density regression [6] under a conditional variable, but the conditional variable is restricted to the multivariate coordinates of samples in their space. Critically, this space can be geographic, spatio-temporal, or any other kind. It can have a dimension of any positive integer. Assuming these dimensions are independent, the DLGM framework calculates a weight for each according to their own bandwidth and kernel function. The product of these weights is used as the final weight to calibrate the least-squared model at each location. This modelling method can be easily adapted to any data of coordinates without trying to collapse the multiple dimensions into a single distance metric in the first instance. Simulation experiments demonstrate that this method is flexible, extensible and customisable. It can also reach higher goodness of fit than specially designed GWR-like models that attempt to accommodate spaces and coordinate systems that are not solely geographical.

2 Methodology

Geographically weighted regression can be expressed as Equation 1 for the sample i at location \mathbf{u}_i ,

$$y_i = \beta_{0i}(\mathbf{u}_i) + \beta_{1i}(\mathbf{u}_i)x_{1i} + \beta_{2i}(\mathbf{u}_i)x_{2i} + \cdots + \beta_{pi}(\mathbf{u}_i)x_{pi} + \epsilon_i \quad (1)$$

and the estimator for its coefficients $\beta_i = (\beta_{0i}, \beta_{2i}, \cdots, \beta_{pi})$ is shown in Equation 2,

$$\hat{\beta}_i = (\mathbf{X}^T \mathbf{W}_i \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W}_i \mathbf{y} \quad (2)$$

where $\mathbf{y} = (y_1, y_2, \cdots, y_n)^T$ is the vector of dependent variables, n is the number of samples, \mathbf{X} is the design matrix or independent matrix of all independent variables, $\epsilon_i \sim N(0, \sigma^2)$ is the random error and \mathbf{W}_i is the geographical weighting matrix for this sample. This weighting matrix is a $n \times n$ diagonal matrix. Each diagonal element is a distance-decay weight $w_{ij} = k(d_{ij}; b)$ (for $j = 1, 2, \cdots, n$) in which d_{ij} is the distance from sample i to j , k is a kernel function and b is the bandwidth. The basic GWR model uses straight-line distance, Minkowski distance, network distance, or travel time [4], which are all spatial. The GTWR model uses the spatial-temporal distance $d_{ij}^{ST} = d_{ij}^S \oplus d_{ij}^T$ by combining spatial distance and temporal distance together [2]. The bandwidth can be fixed (defined by distance), or adaptive (defined by the number of nearest neighbours).

For DLSM, the weight w_{ij} originates as a product of weights for every dimension in the current space, as shown in Equation 3,

$$w_{ij} = \prod_{h=1}^m w_{ijh} = \prod_{h=1}^m k_h(d_{ijh}; b_h) \quad (3)$$

where m is the number of dimensions in \mathbf{u}_i , k_h is the kernel function for dimension h , b_h is the corresponding bandwidth, $d_{ijh} = |u_{ih} - u_{jh}|$, and u_{ih}, u_{jh} is the coordinates in this dimension of sample i and j . Regardless of whether they are measured as longitude, latitude, height, time, social distance or any other measure of “closeness”, they are all feasible dimensions in this model. The estimator of this model can be that shown in Equation 2 or another locally weighted regression estimator.

The weighting method shown in Equation 3 operationalises multiple values of bandwidths – one for each dimension of the various coordinate spaces. The optimization of these bandwidths uses multidimensional minimisation of a criterion function. Theoretically, any kinds of multidimensional minimizer without derivatives are applicable here. We choose the Nelder-Mead algorithm [5]. The criterion function can be either the cross-validation (CV) value or goodness-of-fit, e.g., AIC function of given bandwidth $\mathbf{b} = (b_1, b_2, \dots, b_m)$, shown in Equation 4 and Equation 5 respectively,

$$\text{CV}(\mathbf{b}) = \sum_{i=1}^n [y_i - \mathbf{x}_i \hat{\boldsymbol{\beta}}_{-1}(\mathbf{b})]^2 \quad \text{or} \quad \text{CV}(\mathbf{b}) = \sum_{i=1}^n |y_i - \mathbf{x}_i \hat{\boldsymbol{\beta}}_{-1}(\mathbf{b})| \quad (4)$$

$$\text{AIC}(\mathbf{b}) = 2n \ln \hat{\sigma} + n \ln 2\pi + n \left[\frac{n + \text{tr}(\mathbf{S})}{n - 2 - \text{tr}(\mathbf{S})} \right] \quad (5)$$

where $\hat{\boldsymbol{\beta}}_{-i}(\mathbf{b})$ is the coefficient estimates for sample i without the sample itself, \mathbf{x}_i is the i -th row of matrix \mathbf{X} , \mathbf{S} is the “hat matrix” in which each row \mathbf{s}_i equals to $\mathbf{x}_i(\mathbf{X}\mathbf{W}_i\mathbf{X})^{-1}\mathbf{X}\mathbf{W}_i$.

3 Experiments

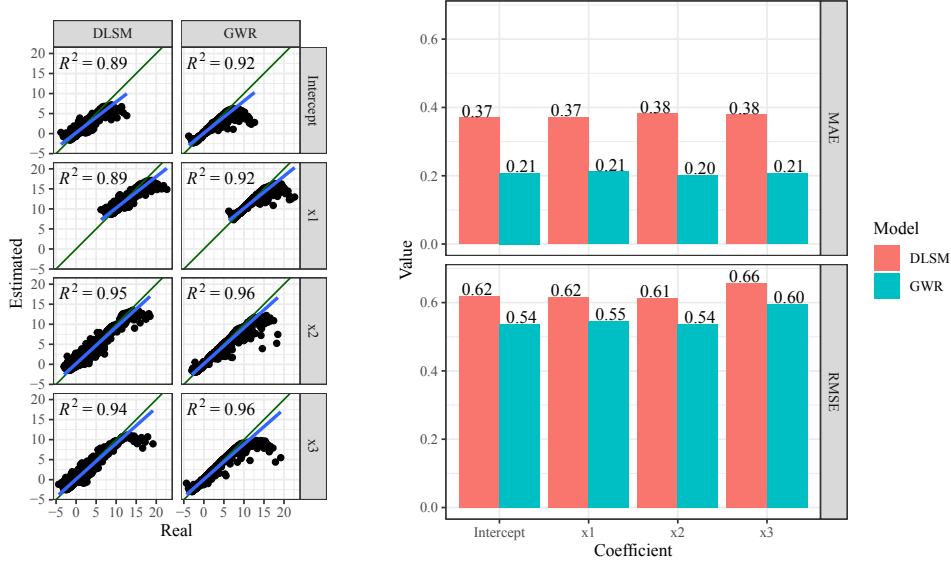
We carried out three experiments, generating simulation data sets to demonstrate how DLSM works². We also calibrated a corresponding GWR-family model in each experiment to provide a comparison. In each experiment, we use root mean squared error (RMSE) or mean absolute error (MAE) to evaluate the precise of estimates, which are defined in Equation 6,

$$\text{RMSE} = \sum_{i=1}^n (r_i - e_i)^2, \quad \text{MAE} = \sum_{i=1}^n |r_i - e_i| \quad (6)$$

where n is the number of estimates, e_i is the i -th estimate, r_i is the corresponding real value.

We first generated a 2D data set of Cartesian coordinates. Anisotropy was preserved in the coefficients. Bandwidths optimized by DLSM are 11.4% (570 neighbours) in the E-W direction and 0.7% (35 neighbours) in the N-S direction. Coefficient estimates and their RMSEs are shown in Figure 1. Whereas DLSM helps identify anisotropy, it is missing in estimates from a basic GWR model because the only bandwidth value optimized by GWR is 16 nearest neighbours (regardless of direction). It also has a stronger risk of overfitting as the bandwidth is too small. By contrast, DLSM can restrain overfitting in dimensions where spatial heterogeneity is weaker.

² Please turn to <https://hpdell.github.io/GIScience-Materials/posts/DLSM/> for more details.



(a) Comparison between estimates and (b) RMSE and MAE of coefficient estimates. real values.

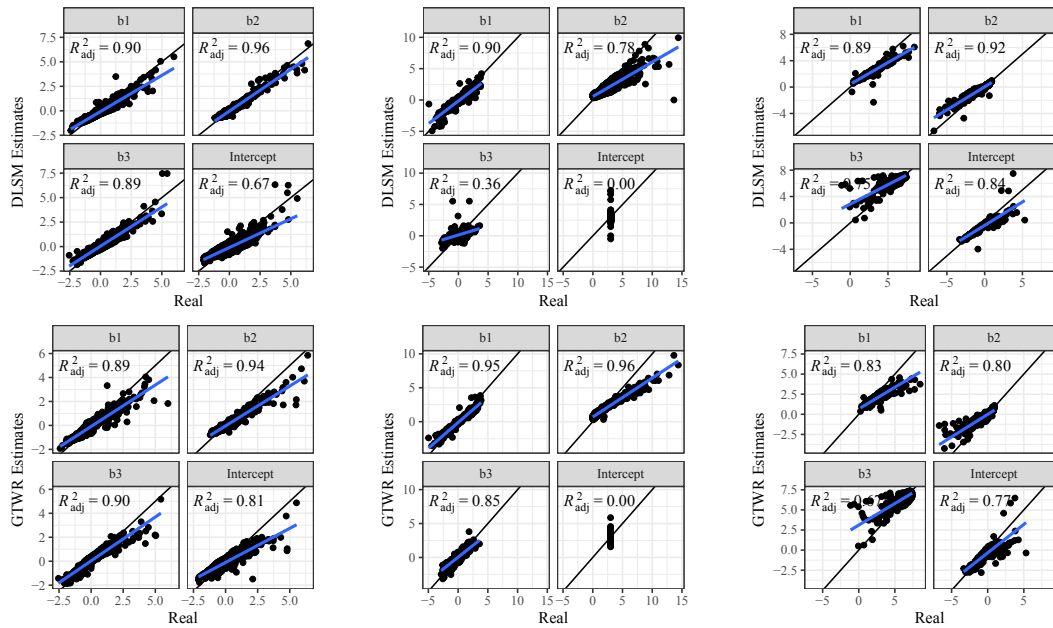
■ **Figure 1** Results of GWDR and basic GWR with two-dimensional spatial data.

Four 3D data sets of Cartesian coordinates representing space-time location (u_1, u_2, u_3) were generated to compare DLSSM and GTWR. In the former two data sets, coefficients were generated by $\exp(u_3)$. While in the latter two data sets, an autoregression model on u_3 was a part of all coefficients. The space-time distance metric use by GTWR was set to $d_{ij}^{ST} = \sqrt{\lambda(\Delta u_{1,ij}^2 + \Delta u_{2,ij}^2) + \mu(\Delta u_{3,ij}^2)}$. Parameters λ and μ in this space-time distance metric were optimized according to goodness of fit. Coefficient estimates and their RMSEs are shown in Figure 2. According to the results, DLSSM can reduce the mean of absolute estimation error by 10%-50%, especially when coefficients are temporally autocorrelated. The multiple bandwidths attach actual meaning to the parameters λ, μ ; they have a real-world correlate, unlike the root of sum of squared meters and seconds ($\sqrt{m^2 + s^2}$).

A 4D data set was also generated to simulate flow data. DLSSM was compared with GWR. For flow data, each flow can be represented by a set of 4D coordinates (u, v, α, l) in which u, v represents the spatial location of its starting point, α represents its direction, and l represents its length. The distance metric used by GWR was set to the similarity between flows $O_i(u_{O_i}, v_{O_i}) \rightarrow D_i(u_{D_i}, v_{D_i})$ and $O_j(u_{O_j}, v_{O_j}) \rightarrow D_j(u_{D_j}, v_{D_j})$ [3], as shown in

$$d_{ij} = \sqrt{\frac{[(u_{O_i} - u_{O_j})^2 + (v_{O_i} - v_{O_j})^2] + [(u_{D_i} - u_{D_j})^2 + (v_{D_i} - v_{D_j})^2]}{l_i l_j}} \quad (7)$$

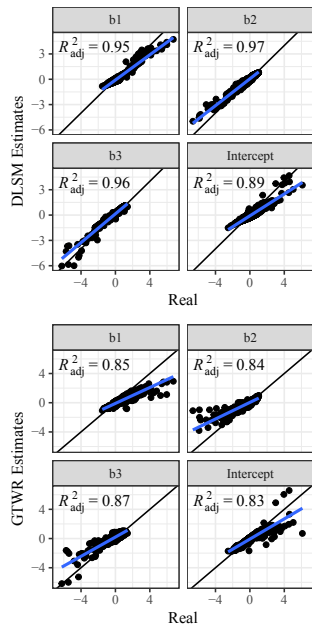
in which l_i is the length of flow $\overrightarrow{O_i D_i}$. Coefficient estimates and their RMSEs are shown in Figure 3. Results show that DLSSM works well for spatial line data even without defining distance metrics. It performs better than GWR according to the mean of estimation errors, but a few outliers exist in estimates. GWR selected a much smaller bandwidth (173 neighbours). Thus, the risk of overfitting reappears.



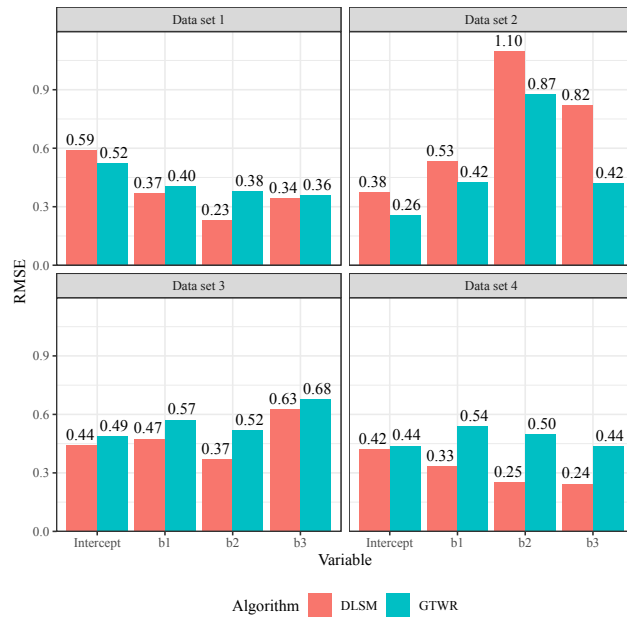
(a) Coefficient estimates and real values, the first data set.

(b) Coefficient estimates and real values, the second data set.

(c) Coefficient estimates and real values, the third data set.

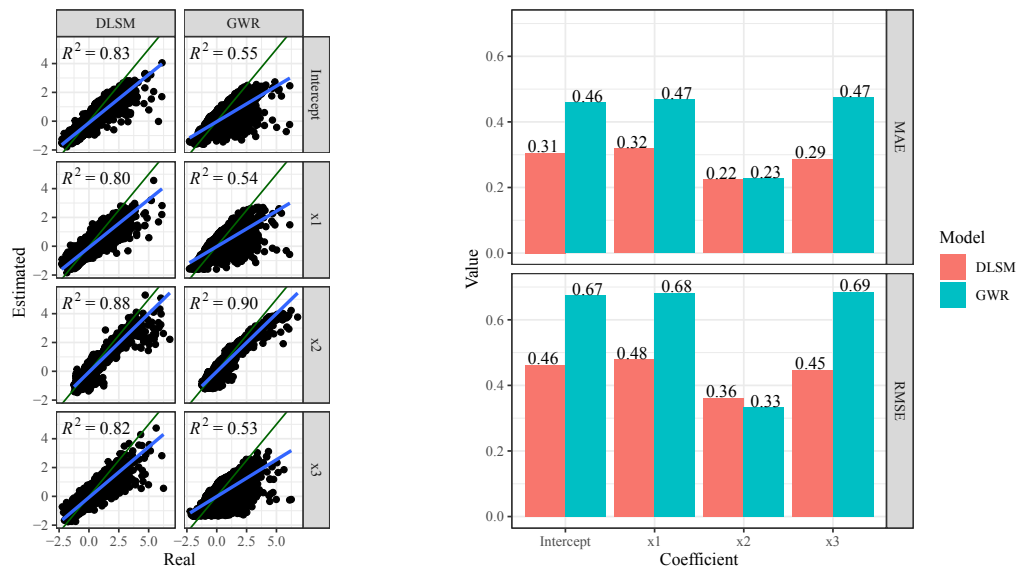


(d) Coefficient estimates and real values, the fourth data set.



(e) RMSE of estimates for each coefficient from DLSM and GTWR models on each data set.

Figure 2 Comparison between real value and estimations of coefficients given by GWDR and GTWR for ordinary spatial and temporal data.



(a) Comparison between estimates and real values. (b) RMSE and MAE of coefficient estimates.

■ **Figure 3** Results of GWDR and basic GWR with four-dimensional spatial data.

4 Conclusion


This paper introduces the DLSTM model as a framework for estimating local regression models, such as GWR and GTWR. It offers more flexibility because of its three alterable parts: a space where samples exist, a set of kernels selected for every dimension and a locally weighted regression method. Simulation shows that DLSTM can be applied to many kinds of spatial data without specially defined distance metrics, such as spatio-temporal data and spatial interaction data. It can also help tackle the effects of anisotropy because it has, in effect, a multidimensional bandwidth and decay function, measuring “closeness” in multiple dimensions simultaneously. In the future, researchers no longer need to design distance metrics to bring together, in a rather ad hoc way, different types of space and coordinate systems into the distance decay function. Assigning a weighting scheme to each of the dimensions and then pooling across them is suggested as a better alternative.

References

- 1 Chris Brunson, A. Stewart Fotheringham, and Martin E. Charlton. Geographically Weighted Regression: A Method for Exploring Spatial Nonstationarity. *Geographical Analysis*, 28(4):281–298, February 1996. doi:10.1111/j.1538-4632.1996.tb00936.x.
- 2 Bo Huang, Bo Wu, and Michael Barry. Geographically and temporally weighted regression for modeling spatio-temporal variation in house prices. *International Journal of Geographical Information Science*, 24(3):383–401, March 2010. doi:10.1080/13658810802672469.
- 3 Maryam Kordi and A. Stewart Fotheringham. Spatially Weighted Interaction Models (SWIM). *Annals of the American Association of Geographers*, 106(5):990–1012, September 2016. doi:10.1080/24694452.2016.1191990.
- 4 Binbin Lu, Martin Charlton, Paul Harris, and A. Stewart Fotheringham. Geographically weighted regression with a non-Euclidean distance metric: A case study using hedonic house

- price data. *International Journal of Geographical Information Science*, 28(4):660–681, April 2014. doi:10.1080/13658816.2013.865739.
- 5 J. A. Nelder and R. Mead. A Simplex Method for Function Minimization. *The Computer Journal*, 7(4):308–313, January 1965. doi:10.1093/comjnl/7.4.308.
- 6 Geoffrey S Watson. Smooth Regression Analysis. *The Indian Journal of Statistics, Series A*, 26(4):359–372, December 1964.

Understanding Place Identity with Generative AI

Kee Moon Jang ✉ 

MIT Senseable City Lab, Cambridge, MA, USA

Junda Chen ✉


DataChat, Madison, WI, USA

Yuhao Kang¹ ✉ 

MIT Senseable City Lab, Cambridge, MA, USA

Junghwan Kim ✉ 

Department of Geography, Virginia Tech, Blacksburg, VA, USA

Jinhyung Lee ✉ 

Department of Geography and Environment, Western University, London, Canada

Fábio Duarte ✉ 

MIT Senseable City Lab, Cambridge, MA, USA

Abstract

Researchers are constantly leveraging new forms of data to understand how people perceive the built environment and the collective place identity of cities. Latest advancements in generative artificial intelligence (AI) models have enabled the creation of realistic representations of real-world settings. In this study, we explore the potential of generative AI as the source of textual and visual information in capturing the place identity of cities assessed by filtered descriptions and images. We asked questions on the place identity of a set of 31 global cities to two generative AI models, ChatGPT and DALL·E2. Since generative AI has raised ethical concerns regarding its trustworthiness, we performed cross-validation to examine whether the results show similar patterns to real urban settings. In particular, we compared the outputs with Wikipedia data for text and images searched from Google for images. Our results indicate that generative AI models have the potential to capture the collective features of cities that can make them distinguishable. This study is among the first attempts to explore the capabilities of generative AI in understanding human perceptions of the built environment. It contributes to urban design literature by discussing future research opportunities and potential limitations.

2012 ACM Subject Classification Social and professional topics → Geographic characteristics

Keywords and phrases ChatGPT, DALL·E2, place identity, generative artificial intelligence, sense of place

Digital Object Identifier 10.4230/LIPIcs.GIScience.2023.41

Category Short Paper

Related Version *Previous Version*: <https://arxiv.org/abs/2306.04662>

1 Introduction

Place identity, often referred to as properties that distinguish a place from others [12, 11], is an important concept in the fields of urban design, geography, tourism, and environmental psychology. As a sense of place that is shaped through diverse human experiences, recognizing such place characteristics has been crucial for understanding human-environment interactions [4, 9, 10]. Yet, measuring and representing place identity has been a challenging task due

¹ Corresponding author: yuhaokang@mit.edu



to the intrinsically subjective nature of place identity. Conventional studies attempted to capture place identity through direct observation, questionnaires, surveys and interviews [4, 10]. In the past decade, researchers have been leveraging new data sources to understand the collective place identity of cities. In particular, two data formats, texts [4, 2, 3], and images such as street-level images and geotagged photos [16, 17] have been effective in revealing place identity information. Urban planners and designers have benefited from these emerging data sources to explore subjective urban experiences and promote data-driven decision-making processes in practices [10].

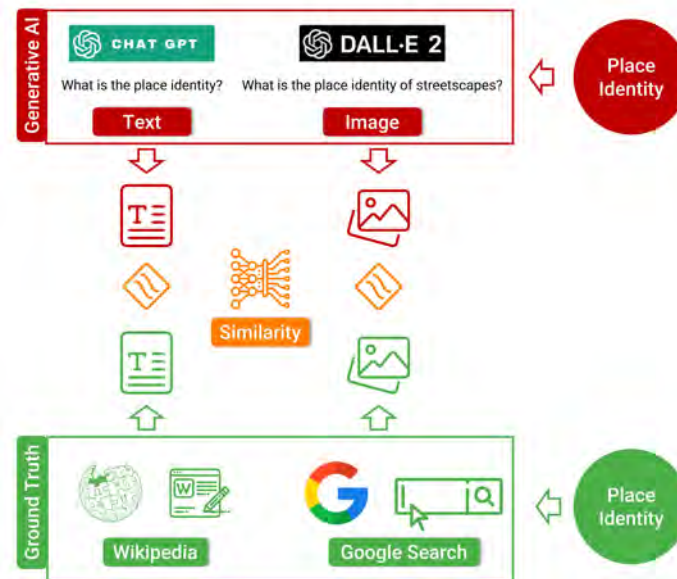
Recently, advancements in generative artificial intelligence (GenAI) models have received significant attention due to their capabilities to generate realistic text and image output based on natural language prompts. ChatGPT and DALL·E2, for instance, have been highlighted as powerful tools with the potential for a wide variety of applications in different domains such as education, transportation, geography, and so forth [6, 7, 8, 14]. Also, there have been attempts in urban studies to evaluate design qualities of the built environment scenes and obtain optimal land-use configuration through automated urban planning process [13, 15]. Despite its promise in urban science, the use of generative AI also faces common ethical concerns such as misinformation and bias, falling short in depicting composition and locales for specific conditions [5]. Therefore, there remains needs for a more robust quantitative examination and analysis of how well they represent place-specific contexts toward trustworthy outputs in different domains.

To this end, since generative AI models are offering new ways to collect textual and visual information that may represent realistic human responses, this study aims to examine the potential of generative AI as new tools for understanding place identity in different cities. In this endeavor, we address two research questions: (1) Can generative AI models identify place identity of cities? and (2) How reliable are the generated outputs when compared with real-world settings? This study is expected to guide urban researchers in using such tools to generate large volumes of data through a more efficient and cost-effective approach, as well as to study place identity in a data-driven manner, which can facilitate our understanding of urban perception.

2 Methodology

We present a computational framework of this study in Figure 1. The framework involves two datasets that we created to investigate the potential of generative AI models in capturing the place identity of 31 global cities. The first dataset is a text-based dataset that we generated using ChatGPT to understand place identity, using the following prompts: “*What is the place identity of {city}? Give me in ten bullet points*”. To ensure consistency and comparability across different cities included in our dataset, we limited the responses to ten bullet points. By doing so, the generated outputs are concise and structured, and can easily be analyzed and compared. The second dataset is an image-based dataset that we collected using DALL·E2 to generate visual representations of streetscapes in different cities. The prompts used to achieve this are the following: “*What is the place identity of streetscapes of {city}*”? We generated 10 images for every city, where each image has a size of 256*256 pixels. By combining the image-based dataset with the text-based dataset, we aim to provide a comprehensive and multi-modal understanding of the place identity of each city.

We further collected two ground-truth datasets including a text dataset from Wikipedia and an image dataset from Google search. Despite the high performance of generative AI tools in generating realistic outputs, concerns regarding their reliability and accuracy have emerged. Thus, we performed the cross-validation to compare the similarities among



■ **Figure 1** The computational framework of this paper.

these datasets to evaluate whether the results provided by generative AI can be trustworthy. For text similarity, we first segmented the Wikipedia corpus into individual sentences and converted each sentence from both datasets into word embeddings. This was achieved by using a sentence transformer BERT model based on a modified version of MiniLM. Then, we measured cosine similarity for sentence embeddings from ChatGPT responses and Wikipedia corporuses to assess the relevance between the two datasets. We also created word cloud images of each city for a visual comparison between topics covered in ChatGPT and Wikipedia texts.

For image similarity, we measured the Learned Perceptual Image Patch Similarity (LPIPS) [18] to assess the perceptual similarity of images generated by DALL · E2 and collected via Google search. The LPIPS metric evaluates the distance between different image patches and produces scores ranging from 0 to 1, where a lower score indicates greater similarity, and vice versa. Subsequently, we identify the top 3 similar Google images for each DALL · E2-generated image based on similarity scores. These analyses allow us to validate whether the results generated by generative AI models are consistent with the real-world urban settings of each city, providing valuable insights for urban design research and practice.

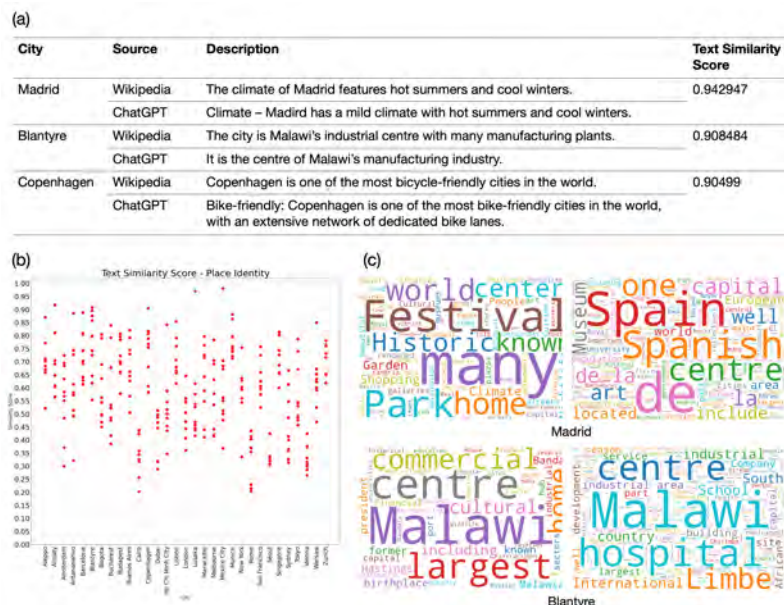
3 Results

3.1 Results of place identity generated by ChatGPT

To validate the accuracy and reliability of the data generated by ChatGPT, we conducted cross-validation with Wikipedia. This involved computing the similarity scores between sentences from ChatGPT and Wikipedia, and presenting visual comparisons between pairs of word clouds. Figure 2 illustrates the validation results. Figure 2(a) shows several examples of high sentence similarity scores. For example, for the city of Madrid, both Wikipedia and ChatGPT-generated sentences had similar descriptions of the climate, resulting in a very high similarity score of 0.94. However, as shown in Figure 2(b), the comparison between

41:4 Understanding Place Identity with Generative AI

ChatGPT-generated sentences and the introduction from Wikipedia resulted in a range of similarity scores, reflecting both similar and dissimilar descriptions of place identity. Such disparities may suggest that there are limitations to the effectiveness of generative AI models in capturing the nuances and complexities of place identity. Last, Figure 2(c) illustrates two cases of word clouds analysis created for ChatGPT responses (left) and Wikipedia (right). While the introduction of Madrid in Wikipedia covered generic keywords such as *spain*, *spanish*, *centre* and *capital*, we found that ChatGPT captures the full spectrum of place identity components as defined in fields of environmental psychology and geography [1, 12]. For instance, topics including *park* and *garden* refer to the “physical settings” of Madrid, *festival* and *shopping* represent the “activities” that take place, and *historic* and *climate* describe the subjective “meanings” that contribute to the place identity formation in the capital of Spain. In the case of Blantyre, the most notable keywords observed in the word cloud of Wikipedia corpus are *Malawi* and *centre*. Likewise, the word cloud generated from ChatGPT response also features the same keywords, from which we infer that ChatGPT identifies the place identity of Blantyre in relation to its significance within the national context of Malawi.

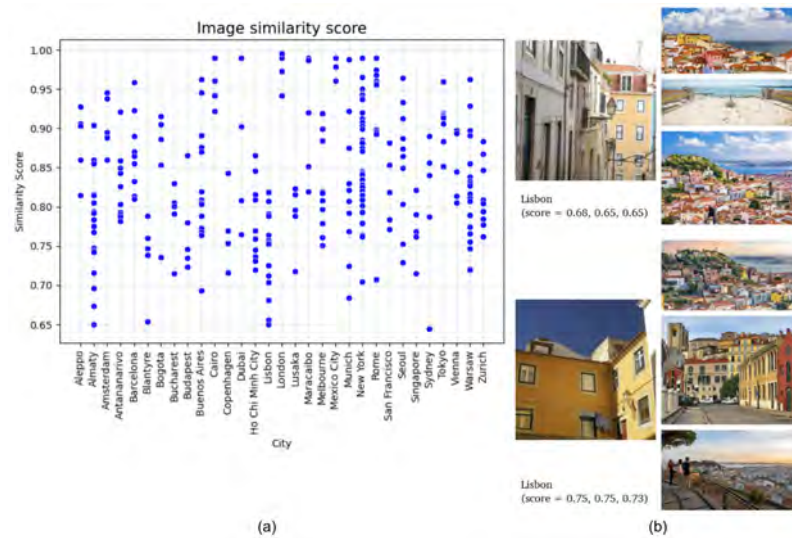


■ **Figure 2** Text similarity results. (a) Examples of high text similarity scores between Wikipedia introductions and ChatGPT responses on place identity; (b) Scatter chart of the distribution of cosine similarity scores between sentences from ChatGPT responses and Wikipedia introduction corpuses; and (c) Word cloud comparison for Madrid and Blantyre cases.

3.2 Results of place and urban identity generated by DALL · E2

Similar to the comparison between ChatGPT-generated sentences with Wikipedia corpus, we also compared images generated by DALL · E2 and those collected from Google search. This was conducted to verify the reliability and generative capability of the text-to-image model in producing realistic representations of place-specific scenes of cities. For this purpose, the image similarity was measured using LPIPS metric to assess the perceptual similarity between AI-generated and real-world images that match well with human judgment. Figure 3

presents the image similarity results. Overall, as shown in Figure 3(a), Almaty, Blantyre, Lisbon and Sydney were cities that reported the highest perceptual similarity with LPIPS value being approximately 0.65. In particular, Lisbon presents relatively consistent low similarity scores within the range of 0.65-0.82. Figure 3(b) shows two examples of DALL · E2 generated images for Lisbon's place identity and their top three matching Google image search results. It is evident that the generative AI effectively captured the low-rise residential buildings with vivid yellow colors in Lisbon, resulting in a low LPIPS score (high similarity). These suggest that, despite variability across cities, DALL · E2 can generate more reliable images of urban scenes for certain cities that reflect their place identity.



■ **Figure 3** Image similarity results. (a) Distribution of LPIPS scores between DALL · E2 generated images and Google image by cities; and (b) Low LPIPS score examples for Lisbon case.

4 Conclusion

In this study, we presented text and image similarity results between responses from two generative AI model, ChatGPT and DALL · E2, and corresponding ground-truth data to test the reliability of their outputs for representing place identity of different cities. Through examining the two datasets, we find that, in many cases, they generated text description or realistic images that represent salient characteristics of cities. In particular, text similarity scores aligned closely with similarities observed in sentence-by-sentence comparison and word clouds of ChatGPT responses and Wikipedia corpuses. This study is among the first to examine the capabilities of generative AI tools in representing the place identity of cities. The overall framework is expected to aid planners and designers in utilizing such tools to identify salient characteristics of cities for sustainable placemaking and city branding purposes.

Despite the contributions of this study, we discuss potential limitations and research opportunities to be addressed in future studies. First, a portion of DALL · E2 generated images is still considered more generic than place-specific, which may not fully reflect the place identity. These images are more relevant to the generic concept of a *city*, rather than *identity*, and fall short in representing the attributes that distinguish a particular city from the rest. Another limitation lies in the uncertainty in the image similarity results. We found that certain similar scenes generated by DALL · E2 resulted in a range of similarity

scores when measured against the same ground-truth image. Yet, it is uncertain why such differences are observed, what contributes to high or low similarity results, and thus which scene is most relevant to the place identity of a particular city.



References

- 1 David Canter. *The psychology of place*. St Martin'S Press, 1977.
- 2 Song Gao, Krzysztof Janowicz, Daniel R Montello, Yingjie Hu, Jiue-An Yang, Grant McKenzie, Yiting Ju, Li Gong, Benjamin Adams, and Bo Yan. A data-synthesis-driven method for detecting and extracting vague cognitive regions. *International Journal of Geographical Information Science*, 31(6):1245–1271, 2017.
- 3 Yingjie Hu, Chengbin Deng, and Zhou Zhou. A semantic and sentiment analysis on on-line neighborhood reviews for understanding the perceptions of people toward their living environments. *Annals of the American Association of Geographers*, 109(4):1052–1073, 2019.
- 4 Kee Moon Jang and Youngchul Kim. Crowd-sourced cognitive mapping: A new way of displaying people's cognitive perception of urban space. *Plos one*, 14(6):e0218590, 2019.
- 5 Yuhao Kang, Qianheng Zhang, and Robert Roth. The ethics of ai-generated maps: A study of dalle 2 and implications for cartography. *arXiv preprint arXiv:2304.10743*, 2023.
- 6 Junghwan Kim and Jinhyung Lee. How does chatgpt introduce transport problems and solutions in north america? *Findings*, 2023.
- 7 Ehsan Latif, Gengchen Mai, Matthew Nyaaba, Xuansheng Wu, Ninghao Liu, Guoyu Lu, Sheng Li, Tianming Liu, and Xiaoming Zhai. Artificial general intelligence (agi) for education. *arXiv preprint arXiv:2304.12479*, 2023.
- 8 Gengchen Mai, Weiming Huang, Jin Sun, Suhang Song, Deepak Mishra, Ninghao Liu, Song Gao, Tianming Liu, Gao Cong, Yingjie Hu, et al. On the opportunities and challenges of foundation models for geospatial artificial intelligence. *arXiv preprint arXiv:2304.06798*, 2023.
- 9 Lynne C Manzo and Douglas D Perkins. Finding common ground: The importance of place attachment to community participation and planning. *Journal of planning literature*, 20(4):335–350, 2006.
- 10 Mahbubur Meenar, Nader Afzalan, and Amir Hajrasouliha. Analyzing lynch's city imageability in the digital age. *Journal of Planning Education and Research*, 42(4):611–623, 2022.
- 11 Harold H Proshansky, Abbe K Fabian, and Robert Kaminoff. Place-identity: Physical world socialization of the self (1983). In *The people, place, and space reader*, pages 111–115. Routledge, 2014.
- 12 Edward Relph. *Place and placelessness*, volume 67. Pion London, 1976.
- 13 Sachith Seneviratne, Damith Senanayake, Sanka Rasnayaka, Rajith Vidanaarachchi, and Jason Thompson. Dalle-urban: Capturing the urban design expertise of large text to image transformers. *arXiv preprint arXiv:2208.04139*, 2022.
- 14 Eva AM van Dis, Johan Bollen, Willem Zuidema, Robert van Rooij, and Claudi L Bockting. Chatgpt: five priorities for research. *Nature*, 614(7947):224–226, 2023.
- 15 Dongjie Wang, Chang-Tien Lu, and Yanjie Fu. Towards automated urban planning: When generative and chatgpt-like ai meets urban planning. *arXiv preprint arXiv:2304.03892*, 2023.
- 16 Fan Zhang, Bolei Zhou, Carlo Ratti, and Yu Liu. Discovering place-informative scenes and objects using social media photos. *Royal Society open science*, 6(3):181375, 2019.
- 17 Fan Zhang, Jinyan Zu, Mingyuan Hu, Di Zhu, Yuhao Kang, Song Gao, Yi Zhang, and Zhou Huang. Uncovering inconspicuous places using social media check-ins and street view images. *Computers, Environment and Urban Systems*, 81:101478, 2020.
- 18 Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

An Integrated Uncertainty and Sensitivity Analysis for Spatial Multicriteria Models

Piotr Jankowski¹  

San Diego State University, CA, USA
Adam Mickiewicz University, Poznan, Poland

Arika Ligmann-Zielińska  

Michigan State University, East Lansing, MI, USA
Adam Mickiewicz University, Poznan, Poland

Zbigniew Zwoliński  

Adam Mickiewicz University, Poznan, Poland

Alicja Najwer  

Adam Mickiewicz University, Poznan, Poland

Abstract

This paper introduces an integrated Uncertainty and Sensitivity Analysis (US-A) approach for Spatial Multicriteria Models (SMM). The US-A approach evaluates uncertainty and sensitivity by considering both criteria values and weights, providing spatially distributed measures. A geodiversity assessment case study demonstrates the application of US-A, identifying influential inputs driving uncertainty in specific areas. The results highlight the importance of considering both criteria values and weights in analyzing model uncertainty. The paper contributes to the literature on spatially-explicit uncertainty and sensitivity analysis by providing a method for analyzing both categories of SMM inputs: evaluation criteria values and weights, and by presenting a novel form of visualizing their sensitivity measures with bivariate maps.

2012 ACM Subject Classification Information systems → Geographic information systems

Keywords and phrases model uncertainty, input factor sensitivity, geodiversity, spatial multicriteria models

Digital Object Identifier 10.4230/LIPICs.GIScience.2023.42

Category Short Paper

Funding This research was supported by the National Science Centre (Narodowe Centrum Nauki) under Grant No. UMO-2018/29/B/ST10/00114.

1 Introduction

Uncertainty analysis (UA) and sensitivity analysis (SA) are two complementary methods of evaluating uncertainty present in model inputs and, by extension, in model results [12]. UA quantifies outcome variability given model input uncertainties, and is, therefore, forward-looking as it focuses on evaluating how the uncertainty of inputs propagates through the model and affects its output values. However, UA does not inform about the magnitude of individual inputs' influence on model output variability. This information can be obtained from SA that relates the output variability to model inputs and evaluates how much each source of uncertainty contributes to the overall variability of the output. In this sense, SA is a backward-looking approach that complements UA.

¹ corresponding author



Spatial Multicriteria Models (SMM) implemented in the context of GIS-based multicriteria analysis employ either value function-based methods or outranking relation-based methods to arrive at a rank-order/classification of spatially-explicit choice alternatives [8]. In SMM that employ value function methods, the rank order is determined by a synthetic score expressing the overall strength of each choice alternative vis-à-vis other alternatives under consideration. The score is calculated by integrating criteria values with weights using a combination rule. Due to potential errors in criteria values and the subjectivity of weights, both types of inputs can become potential sources of uncertainty affecting the SMM output. The overall impact of uncertainty can be represented by a measure of output variability (e.g., variance), which is also a proxy of output uncertainty. In order to isolate influential inputs driving the model's output uncertainty, one can employ SA. Ultimately, the purpose of UA combined with SA is to improve the model's reliability and its value for policy and decision-making.

2 Related work

Two approaches to UA-SA – local and global, have been proposed for SMM. In the local approach, the values of model inputs are varied one at a time (OAT) while keeping other inputs unchanged. This approach has been popular among modelers due to its simplicity, tractability, and low computational cost [15]. Yet, in SMM based on compensatory decision rules (i.e., Weighted Linear Combination, Analytical Hierarchy Process), model inputs do interact, and the OAT approach does not address these interaction effects. In contrast, the global approach accounts for model input interactions by more or less systematically sampling the entire input value space [7]. The downside of the global approach is its computational cost. Different solutions to accelerating global SA for spatial models have been proposed, including parallelization [1], [5] and surrogate models [11].

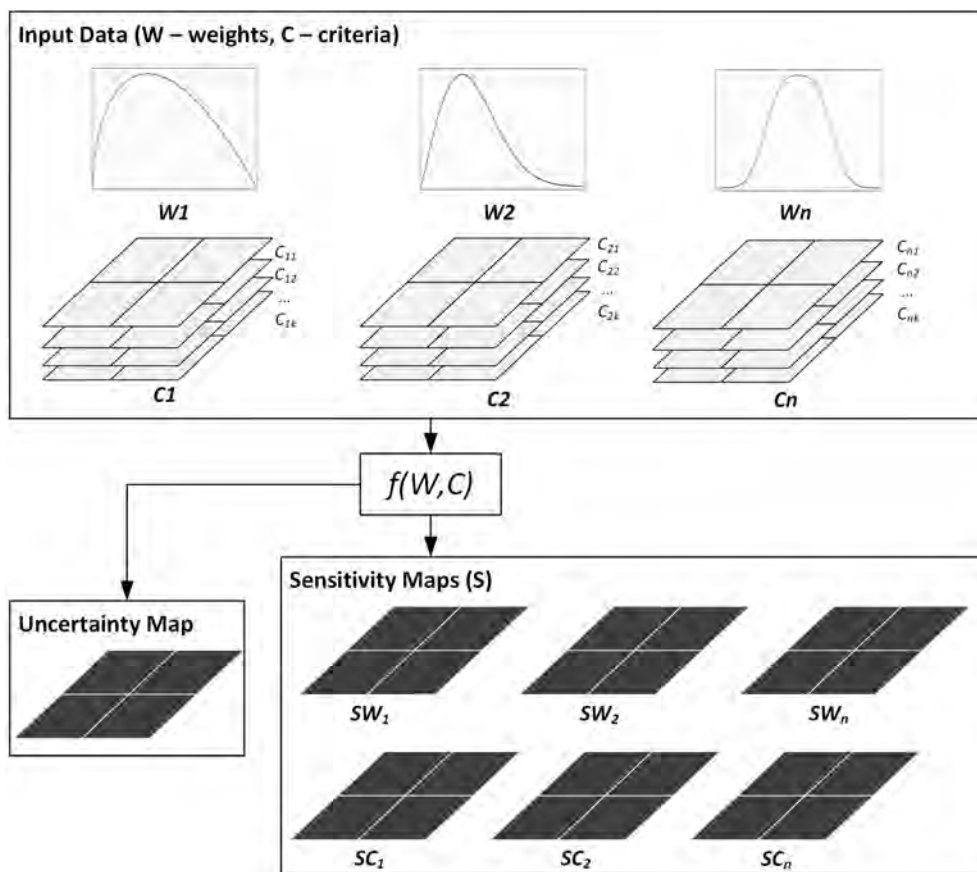
In an early example of global approach for SMM, [3] used variance decomposition-based SA to investigate model's solution stability in light of uncertainties affecting criteria values and weights. In their study, SA was performed on aggregated criteria values and weights, producing one measure of sensitivity for each input for the entire study area. This approach to UA-SA takes spatially explicit inputs, identifies among them the influential ones that drive the model's output variability, and returns non-spatial estimates of sensitivity without providing a crucial piece of information – namely, where in the study area this influence plays out. Others, including [6], [2], and [10] proposed a spatially explicit and integrated approach to UA-SA of SMM, henceforth referred to as US-A, based on global variance decomposition, in which the output of SMM results in spatially distributed measures of uncertainty and sensitivity. Their work, however, addressed only one category of uncertain inputs: criteria weights. The work presented here extends it by providing a method for analyzing both categories of SMM input: criteria values, and weights. Additionally, it presents a novel form of visualizing their sensitivity measures with bivariate maps.

3 Methods

The US-A of variable criteria and weights is presented in Figure 1. In this approach, weights W_n are represented as probability distributions, whereas criteria C_n are represented as sets of k multiple alternative layers. Since both types of inputs are stochastic, a given SMM $f(W, C)$ has to be calculated multiple times, each time with a different vector of input values. Each calculation uses n scalars for W and n maps for C , where the scalars are derived from weights' respective probability distributions and the maps from their respective sets of realizations.

The sampling used to generate the vectors is called Sobol’s quasi-random with radials and is described in [14]. As a result, we obtain a distribution of SMM spatial outputs, for which we can calculate different aggregation statistics’ maps like mean or standard deviation. Both statistics can then be used jointly (Figure 2) as an uncertainty map.

The next step involves spatially-explicit variance decomposition, independently applied to every spatial unit (*su*) in the study area (e.g., raster cell, vector polygon). Variance decomposition involves subdividing the total variance of *su* creating partial variances for each input [14], [13]. The procedure produces two sensitivity indices per input – First Order Effects Index and Total Effects Index. The former is the input’s fractional contribution to the total variance when the given input is treated independently from all other inputs. The latter is the input’s fractional contribution to the total variance due to its independent influence and interactions with other inputs. Consequently, the difference between the Total Effects Index and the First Order Effects Index is the input’s interactions (Figure 3, legend). The final results comprise 2N sensitivity maps (i.e., one map per each W and one per each C) depicting regions of input’s combined (i.e., bivariate) “first order and interactions” influence on SMM outputs.



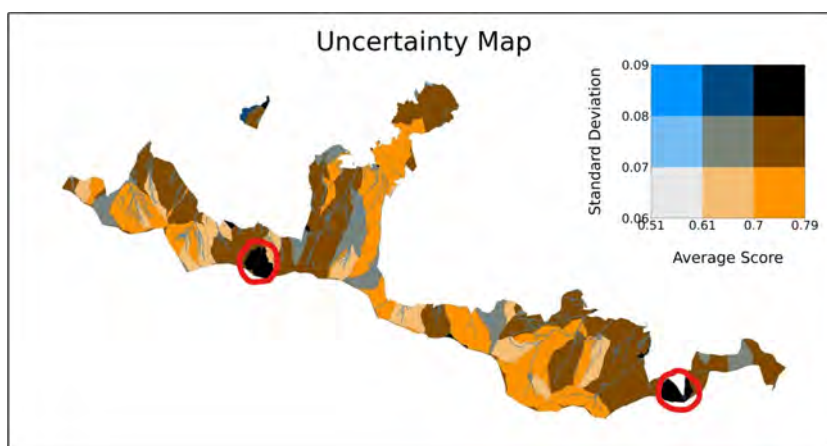
■ **Figure 1** A framework for an extended US-A incorporating the analysis of criteria values and weights.

4 Case study

US-A was employed to assess the uncertainty and sensitivity of multicriteria geodiversity assessment [16], for the Karkonosze National Park (KNP) in southwestern Poland. The park is known for its unique relief and the richness of landforms, including mountain-top planation surfaces, glacial kettles, granite tors with fanciful shapes, waterfalls, and peat bogs. A multicriteria model developed for the purpose of assessment included seven criteria (lithological features, relief energy, landforms, land cover and land use, soils, solar radiation and the topographical wetness index), their relative importance weights, and it was based on a weighted linear combination function for aggregating criteria values with weights. The criteria values and weights were collected from 57 experts in geodiversity and/or Earth sciences using a geo-questionnaire [4]. The study area, the model, and the data collection approach are described in detail in [9].

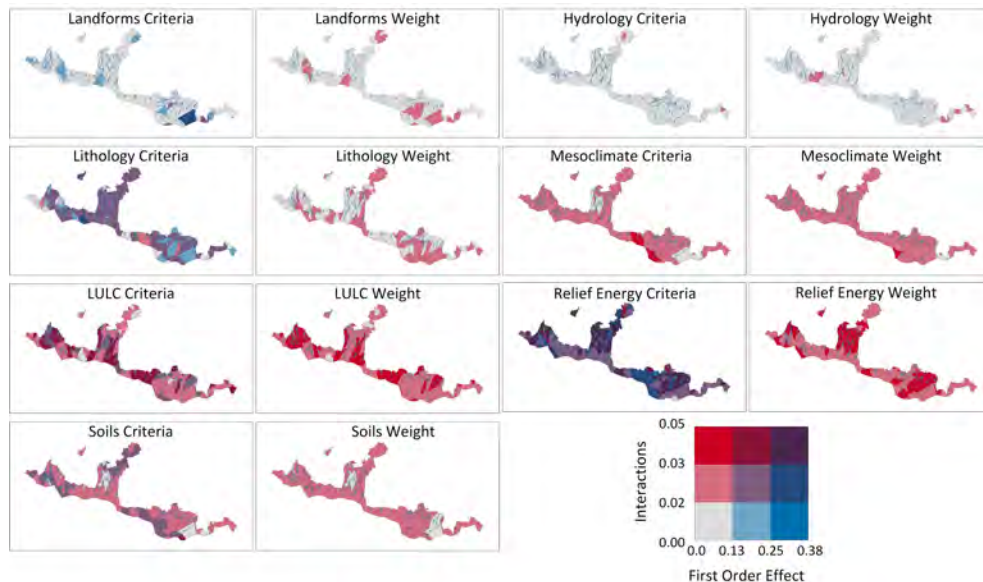
5 Results

Uncertainty analysis is the first step of US-A (Fig. 1). Figure 2 shows its results, including 1) standardized, average geodiversity score (0.0 – 1.0 scale) calculated for each of 212 first order watersheds (assessment units) based on 2000 model runs, and 2) standard deviation representing the measure of uncertainty. Each model run used a sample of input values drawn from discrete uniform probability distributions of criteria maps discrete non-uniform distributions of weights. The sampling scheme was based on Sobol's quasi-random sampling sequence that improves the uniformity of samples in the parameter space [13]. Many watersheds in Fig. 2 exhibit high average values of geodiversity (0.79 – 0.7) and medium-low standard deviation (0.08 – 0.06). We focus our analysis on three watersheds rendered in black in Fig. 2, representing high average geodiversity (0.79 – 0.7) and relative high uncertainty (0.09 – 0.08). These watersheds, which are highlighted in red circles (Fig. 2), represent areas characterized by the richness of geomorphological forms. Two of them (lower right red circle), located in the eastern part of the park, include the headwaters of Sowie Valley in the eastern part of Black Range. The third watershed, located in the western part of the park (upper right), covers Snow Kettles – the second deepest complex of glacial kettles in the park.



■ **Figure 2** Spatial distribution of average geodiversity and standard deviation in KNP.

In order to identify inputs driving the uncertainty of the selected watersheds, we used the combined “first order and interactions” effects for each of the model’s 14 inputs (seven criteria + seven weights). As described in section 3, variance decomposition produces two sensitivity indices for each criterion and each weight. A challenge in mapping first and total effects sensitivity indexes in the presence of many inputs is cognitive difficulty in interpreting 2N sensitivity maps. The values of indexes are typically rendered on coincident maps (side-by-side) requiring a lot of visual back and forth. To overcome this challenge, we used a bivariate map for each input, which allowed us to present the distribution of both index values on one map per input (Fig. 3). The examination of the sensitivity maps in Figure 3 reveals that both landforms and lithology criteria contribute to a relatively high uncertainty (high standard deviation) of geodiversity values in the three watersheds. Specifically, the landforms criterion affects geodiversity of the watersheds covering Sovia Valley and the eastern part of Black Range (lower right) and the lithology criterion impacts geodiversity of the watershed covering Snow Kettles (upper right). This could be addressed, for example, by obtaining higher quality input data for the criteria, which in turn might reduce the uncertainty of assessment. The other input contributing to high uncertainty is the relief energy criterion, but only for the watershed that covers Snow Kettles (upper left).



■ **Figure 3** Spatial distribution of First Order and Interactions (Total Order) effects across 14 input factors.


6 Conclusion

The work presented here shows that considering only criteria weights in US-A may give us an incomplete understanding of important factors driving multicriteria model output uncertainty. Notably, the framework presented in Figure 1 lends itself to incorporating in US-A potential sources of the model’s output uncertainty other than criteria values and weights. Other considerations, not accounted for in this study, are the model’s decision rule represented by aggregation function(s) and the selection of criteria used in the model. They can be addressed in future research.

References

- 1 Christoph Erlacher, Karl-Heinrich Anders, Piotr Jankowski, Gernot Paulus, and Thomas Blaschke. A framework for cloud-based spatially-explicit uncertainty and sensitivity analysis in spatial multi-criteria models. *ISPRS Int. J. Geo Inf.*, 10(4):244, 2021. doi:10.3390/ijgi10040244.
- 2 Bakhtiar Feizizadeh, Piotr Jankowski, and Thomas Blaschke. A GIS based spatially-explicit sensitivity and uncertainty analysis approach for multi-criteria decision analysis. *Comput. Geosci.*, 64:81–95, 2014. doi:10.1016/j.cageo.2013.11.009.
- 3 M. Gómez-Delgado and Stefano Tarantola. GLOBAL sensitivity analysis, GIS and multi-criteria evaluation for a sustainable planning of a hazardous waste disposal site in Spain. *Int. J. Geogr. Inf. Sci.*, 20(4):449–466, 2006. doi:10.1080/13658810600607709.
- 4 Piotr Jankowski, Michal Czepkiewicz, Marek Mlodkowski, and Zbigniew Zwolinski. Geo-questionnaire: A method and tool for public preference elicitation in land use planning. *Trans. GIS*, 20(6):903–924, 2016. doi:10.1111/tgis.12191.
- 5 Jeon-Young Kang, Alexander Michels, Andrew Crooks, Jared Aldstadt, and Shaowen Wang. An integrated framework of global sensitivity analysis and calibration for spatially explicit agent-based models. *Trans. GIS*, 26(1):100–128, 2022. doi:10.1111/tgis.12837.
- 6 Arika Ligmann-Zielinska and Piotr Jankowski. Spatially-explicit integrated uncertainty and sensitivity analysis of criteria weights in multicriteria land suitability evaluation. *Environ. Model. Softw.*, 57:235–247, 2014. doi:10.1016/j.envsoft.2014.03.007.
- 7 Linda Lilburne and Stefano Tarantola. Sensitivity analysis of spatial models. *Int. J. Geogr. Inf. Sci.*, 23(2):151–168, 2009. URL: <http://www.informaworld.com/smpp/content%7Edb=all%7Econtent=a902651821%7Efrm=abslink>.
- 8 Jacek Malczewski and Piotr Jankowski. Emerging trends and research frontiers in spatial multicriteria analysis. *Int. J. Geogr. Inf. Sci.*, 34(7):1257–1282, 2020. doi:10.1080/13658816.2020.1712403.
- 9 Alicja Najwer, Piotr Jankowski, Jacek Niesterowicz, and Zbigniew Zwolinski. Geodiversity assessment with global and local spatial multicriteria analysis. *Int. J. Appl. Earth Obs. Geoinformation*, 107:102665, 2022. doi:10.1016/j.jag.2021.102665.
- 10 Seda Salap-Ayça and Piotr Jankowski. Integrating local multi-criteria evaluation with spatially explicit uncertainty-sensitivity analysis. *Spatial Cogn. Comput.*, 16(2):106–132, 2016. doi:10.1080/13875868.2015.1137578.
- 11 Seda Salap-Ayça, Piotr Jankowski, Keith C. Clarke, Phaedon C. Kyriakidis, and Atsushi Nara. A meta-modeling approach for spatio-temporal uncertainty and sensitivity analysis: an application for a cellular automata-based urban growth and land-use change model. *Int. J. Geogr. Inf. Sci.*, 32(4):637–662, 2018. doi:10.1080/13658816.2017.1406944.
- 12 Andrea Saltelli and Paola Annoni. How to avoid a perfunctory sensitivity analysis. *Environ. Model. Softw.*, 25(12):1508–1517, 2010. doi:10.1016/j.envsoft.2010.04.012.
- 13 Andrea Saltelli, Paola Annoni, Ivano Azzini, Francesca Campolongo, Marco Ratto, and Stefano Tarantola. Variance based sensitivity analysis of model output. design and estimator for the total sensitivity index. *Comput. Phys. Commun.*, 181(2):259–270, 2010. doi:10.1016/j.cpc.2009.09.018.
- 14 Ilya M. Sobol. Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. *Mathematics and Computers in Simulation*, 55(1-3):271–280, 2001. doi:10.1016/S0378-4754(00)00270-6.
- 15 Chen Yun, Jia Yu, and Shahbaz Khan. The spatial framework for weight sensitivity analysis in AHP-based multi-criteria decision making. *Environmental Modelling and Software*, 48(October 2013):129–140, 2013. doi:10.1016/j.envsoft.2013.06.010.
- 16 Zbigniew Zwoliński, Alicja Najwer, and Marco Giardino. *Geoheritage: Assessment, Protection, and Management*, chapter 2, pages 27–52. Elsevier, Amsterdam, The Netherlands, 2018.

Evaluating the Effectiveness of Large Language Models in Representing Textual Descriptions of Geometry and Spatial Relations

Yuhan Ji ✉ 

GeoDS Lab, Department of Geography, University of Wisconsin-Madison, WI, USA

Song Gao ✉ 

GeoDS Lab, Department of Geography, University of Wisconsin-Madison, WI, USA

Abstract

This research focuses on assessing the ability of large language models (LLMs) in representing geometries and their spatial relations. We utilize LLMs including GPT-2 and BERT to encode the well-known text (WKT) format of geometries and then feed their embeddings into classifiers and regressors to evaluate the effectiveness of the LLMs-generated embeddings for geometric attributes. The experiments demonstrate that while the LLMs-generated embeddings can preserve geometry types and capture some spatial relations (up to 73% accuracy), challenges remain in estimating numeric values and retrieving spatially related objects. This research highlights the need for improvement in terms of capturing the nuances and complexities of the underlying geospatial data and integrating domain knowledge to support various GeoAI applications using foundation models.

2012 ACM Subject Classification Computing methodologies → Artificial intelligence

Keywords and phrases LLMs, foundation models, GeoAI

Digital Object Identifier 10.4230/LIPIcs.GIScience.2023.43

Category Short Paper

Funding The authors would like to acknowledge the support from the H.I. Romnes Fellowship, National Science Foundation (No. 2112606) and Arity.

1 Introduction

Deep learning methods have exhibited great performance to tackle many challenging tasks in geographical sciences [16, 9]. However, the models often depend on handcrafted features for specific downstream tasks, thus being hard to be generalized into different tasks. The emergence of representation learning largely mitigated the issue by decomposing the learning process into two steps (task-agnostic data representation and downstream task) [1]. Therefore, an effective location-based representation should preserve key spatial information (e.g., distance, direction, and spatial relations) and make classifiers or other predictors easy to extract useful knowledge [13]. In geospatial artificial intelligence (GeoAI) research, although the geospatial data are usually well-formatted and can be readily understood by GIS software, not all of them can be directly integrated into a deep learning model.

The success of ChatGPT has been a milestone that attracts the general public's attention to Large Language Models (LLMs). With tons of parameters trained on a large text corpus, LLMs have learned profound knowledge across many domains. Other well-known LLMs include the Bidirectional Encoder Representations from Transformers (BERT) [5], the Generative Pre-trained Transformer (GPT) series [14, 2], etc. Despite the differences in network architectures, these LLMs can achieve state-of-the-art performance on natural language processing (NLP) benchmarks. Consequently, researchers have begun the early exploration of integrating LLMs into GIS research, such as geospatial semantic tasks [12] and



© Yuhan Ji and Song Gao;

licensed under Creative Commons License CC-BY 4.0

12th International Conference on Geographic Information Science (GIScience 2023).

Editors: Roger Beecham, Jed A. Long, Dianna Smith, Qunshan Zhao, and Sarah Wise; Article No. 43; pp. 43:1–43:6

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

automating spatial analysis workflows [11]. These studies have demonstrated the ability of LLMs to understand and reason about geospatial phenomena from a semantic perspective as learned from human discourse or formalized programming instructions. In contrast, accurate geometries and spatial relations in GIS are not necessarily expressed in natural languages. Therefore, it can be challenging for LLMs to reconstruct the physical world solely from the textual description of these building blocks, which is the motivation of this research.

In GIScience, spatial relations refer to the connection between spatial objects regarding their geometric properties [8], which play an important role in spatial query, reasoning and question-answering. Using natural language to describe spatial relations is essential for humans to perceive our surroundings and navigate through space. Attempts have been made to formalize the conversion between quantitative models and qualitative human discourse [4]. For topological spatial relations, the RCC-8 (region connection calculus [15]) and the Dimensionally Extended 9-intersection (DE-9IM) model [6] are widely used. Based on the DE-9IM model, five predicates are named by [3] for complex geometries, including *crosses*, *disjoint*, *touches*, *overlaps*, *within*. On top of them, the Open Geospatial Consortium (OGC) further added the predicates *equals*, *contains*, *intersects* for computation convenience. In addition, predicates can also be used to describe the distance or direction between a subject and an object. Fuzzy logic can also be adopted to convert precise metrics into narrative predicates such as *near* and *far* [18].

However, there remains a gap between the contextual semantics of predicates in everyday language and the abovementioned formalization procedures, yielding disagreement and vagueness in the understanding. It is yet to be determined whether the LLMs can fully capture how people describe spatial objects with predicates in natural language. If so, how we can leverage such knowledge to represent geospatial contexts with LLMs.

2 Methodology

2.1 Workflow

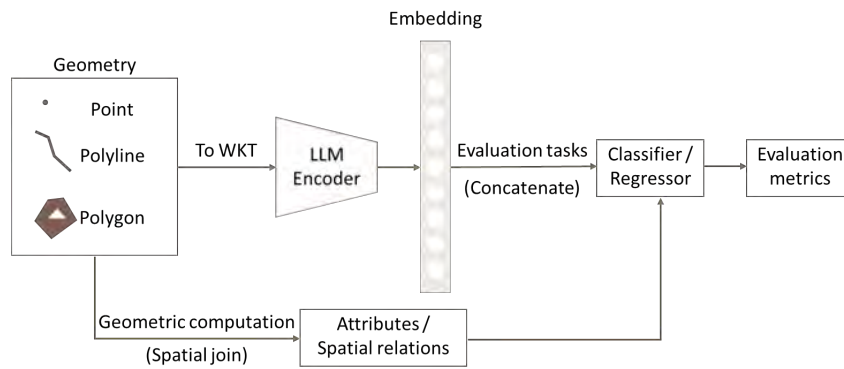
This research focuses on assessing the ability of LLMs in representing geometries and their spatial relations through a set of downstream tasks. Figure 1 illustrates the workflow we employed, which consists of three primary modules. The first module utilizes a GIS tool to extract the attributes, such as geometry type, centroid, and area, of individual geometries and their spatial relations, including predicates and distances between pairs of geometries. The second module applies LLMs to encode the well-known text (WKT) format of geometries, e.g., `LINESTRING (30 10, 10 30, 40 40)`, which includes the geometry type and the ordered coordinates whereas the map projection is not considered in this work. Finally, the obtained embeddings from LLMs, along with the ground-truth attributes or spatial relations, are fed into classifiers or regressors to evaluate the effectiveness of the LLMs-based embeddings.

2.2 Notation

The notations used in this paper are listed in Table 1.

2.3 Evaluation Tasks

The downstream tasks are designed for deriving the geometric attributes or identifying spatial relations, as described in Table 2. The targets of Tasks 1-5 are straightforward, that is, to train a neural network classification/regression model that can best approximate the ground-truth values computed from a GIS tool. All of these tasks use a Multilayer Perceptron



■ **Figure 1** The evaluation workflow of this research.

■ **Table 1** Notations.

Notation	Description
g	A geometry instance (e.g. Point, LineString, and Polygon) that can be processed in GIS tools
$WKT(g)$	The WKT format of g
$Enc(g)$	The location encoding of g using a LLM model to encode $WKT(g)$
$Type(g)$	The geometry type of g
$Centroid(g)$	The centroid of g
$Area(g)$	The area of g
rel	A predicate that can be used to represent the spatial relation, which is one of {equals, disjoint, intersects, crosses, touches, contains, within, overlaps}, as defined by OGC and implemented in GeoPandas.
$Rel(g_i, g_j)$	The spatial relation between the subject g_i and the object g_j
$Dist(g_i, g_j)$	The minimum euclidean distance between two objects g_i and g_j
$[Enc(g_i); Enc(g_j)]$	The concatenation of the embeddings of g_i and g_j
$Enc(rel, g)$	The embedding of the short phrase $rel + WKT(g)$. For example, “within Polygon ((0 0, 0 1, 1 1, 1 0, 0 0))”

(MLP) as the classifier or regressor. Task 6 aims to investigate whether a geometry g_i can be predicted based on its neighbor g_j and their spatial relation $Rel(g_i, g_j)$. We employ the nearest neighbor retrieval approach to evaluate whether LLMs have learned the meaning of spatial predicates properly. During inference, given an object g_j and a spatial relation rel , we retrieve the top-k nearest neighbors of $Enc(rel, g_j)$ and examined whether they belong to the set of subjects $\{g_i | Rel(g_i, g_j) = rel\}$. This approach assesses the ability of the LLMs to relate geographic objects through spatial predicates.

■ **Table 2** Evaluation Tasks.

Task	Subtask	Model type	Input	Target
Geometric attributes	T1: Geometry type	Classification	$Enc(g)$	$Type(g)$
	T2: Area computation	Regression	$Enc(g)$	$Area(g)$
	T3: Centroid derivation	Regression	$Enc(g)$	$Centroid(g)$
Spatial relations	T4: Spatial predicate	Classification	$[Enc(g_i); En(g_j)]$	$Rel(g_i, g_j)$
	T5: Distance measure	Regression	$[Enc(g_i); En(g_j)]$	$Dist(g_i, g_j)$
	T6: Location prediction	Retrieval	$Enc(rel, g_j)$	$\{g_i Rel(g_i, g_j) = rel\}$

3 Experiments

3.1 Dataset and Preprocessing

Since there is no available benchmark dataset, we constructed real-world multi-sourced geospatial datasets for our case study in Madison, Wisconsin, United States. We downloaded the OpenStreetMap road network data (including links and intersections) using *OSMnx*¹, points of interest (POIs) categorized by *SLIPO*², and *Microsoft Building Footprints*³. Our evaluation tasks focus on the spatial objects with *Point*, *LineString*, and *Polygon* geometry types and assessing their spatial relations, respectively. The datasets are created as follows.

1) For each geometry type, we randomly select 4,000 samples, including 2,000 road intersections and 2,000 POIs for *Point* data, 4,000 road links for *LineString* data, and 4,000 building footprints for *Polygon* data. In total 12,000 samples are used for performing the downstream tasks. The area and centroid of each polygon are also computed.

2) For the spatial predicate *disjoint*, we randomly generate pairs of geometries and check whether their spatial relation is disjoint. For other predicates, we identify spatially related objects using spatial join. Given each combination of subject/object geometry type and their spatial predicate, we keep 400 triplets (subject, predicate, object) for each category for the task of predicate prediction and distance measure. Then we compute the minimum distance between the subjects and the objects.

3) We further construct data for the task of location prediction. In addition to the subjects and objects that are spatially joined in step 2), we also relate neighboring disjoint geometries using a buffer radius of 0.003°. The predicate of “disjoint” is replaced by “disjoint but near”. For each predicate except *disjoint*, we select 200 objects of each geometry type that are related to more than 5 subjects by the same predicate.

All the computations are performed by using the *GeoPandas* package in Python. We consider the predicates of *crosses*, *disjoint (but near)*, *touches*, *overlaps*, *within*, *equals*, *contains* in this work but not *intersects* as it is the opposite of *disjoint*. The data for downstream tasks are further split into 80% training, 5% validation, and 15% test sets.

3.2 Encoding

In this work, we perform the evaluation tasks based on two LLMs: GPT-2 and BERT. Due to the computational and memory resources required to train and use the models, GPT-2 and BERT have a maximum input sequence length (i.e., 1024 and 512 tokens respectively). Therefore, a sliding window approach is employed to tackle the issue as the WKT of *LineString* and *Polygon* types can exceed the length limitation. The long input sequences are broken down into smaller segments of 512 tokens with an overlap of 256 tokens between adjacent segments. Each segment is processed by the LLMs separately. We then take the average of the token embeddings to generate the final embedding for the whole sequence of geometries.

3.3 Training MLPs

As we hypothesize that the learned embeddings from LLMs can be effectively utilized in downstream geometry-related tasks, we use a simple neural network architecture (i.e., MLP) across all tasks. Specifically, the input layer of the MLP is the embedding layer generated from LLMs, followed by a dropout layer for regularization purposes. Following the dropout

¹ <http://osmnx.readthedocs.io/>

² <http://slipo.eu/>

³ <http://www.microsoft.com/maps/building-footprints>

layer is a single hidden layer, which employs the Rectified Linear Unit (ReLU) activation function. Finally, the MLP is concluded with the output linear layer. The number of neurons in the output layer varies depending on the specific task.

To facilitate the training process, we apply a logarithmic function to the target values for the area computation and distance measure tasks. In the centroid derivation task, we use the min-max normalization for the target values. The loss function combines the Mean Squared Error (MSE) on both the transformed and original scales. However, for reporting the performance, we only use the original scale of the target values.

3.4 Results

As shown in Table 3, the performance of the downstream tasks based on the embeddings generated by GPT-2 and BERT are similar, which can be understood from the similarity in their subword tokenization and transformer-based architecture.

■ **Table 3** LLMs Performance Comparison.

Tasks		Metric	GPT-2		BERT	
			Validation	Test	Validation	Test
T1: Geometry type		Accuracy(%)	100	100	100	100
T2: Area computation	All geometries	MAPE(%)	13124	11700	12251	10850
	Polygon only		45.1	44.1	40.7	41.9
T3: Centroid derivation		RMSE	0.037	0.037	0.029	0.029
T4: Spatial predicate	Without geometry type	Accuracy(%)	62.6	65.7	63.8	68.7
	With geometry type		73.7	71.0	73.1	72.3
T5: Distance measure	Disjoint only	RMSE	0.064	0.063	0.057	0.075
T6: Location prediction		Precision@5	N/A	0.03	N/A	0.03

For T1-T3, the assessment is conducted on individual geometries. The 100% accuracy achieved on both the validation and the test dataset of T1 is expected as the geometry type are words that often occur in text documents. Considering the unit of *degree* in longitude and latitude, significant errors (measured by Mean Absolute Percentage Error (MAPE) and Root Mean Square Error (RMSE)) are observed in area and centroid computations, and increasing or reducing the model complexity does not alleviate the issue, suggesting a potential loss of information when averaging the token embeddings or fragmentation of coordinates during tokenization. Training the regressor on all geometries for T2 does not successfully learn that *Point* and *LineString* have an area of 0. Even when training the regressor on *Polygon* separately, the results remain unsatisfactory. In T3, the centroids computed from the high-dimensional embeddings often fall outside the study area. T4-T6 evaluates the embeddings' ability to capture spatial relations. One interesting finding is that the spatial predicate can be better predicted when combined with the geometry type, with accuracy increased from 62%~68% to 71%~73%. This can be attributed to the imbalanced spatial relations among different combinations of geometry types. However, the distance measure task T5 still faces challenges in accurately estimating numeric values even when restricted to the "disjoint" relation only. The poor performance on T6 shows that even though the LLMs can encode the spatial relations and geometries in a consistent way, generating embeddings using an average approach alone is insufficient to support spatial reasoning and conduct geometric manipulations directly. Therefore, a different design to enhance the function of localizing spatial objects from textual descriptions [17] can improve the applications of LLMs in GeoAI.

Overall, the results indicate that the LLMs-generated embeddings have encoded the geometry types and coordinates present in the WKT format of geometries. However, it should be noted that the performance of the embeddings does not consistently meet expectations across all evaluation tasks. While the LLMs-generated embeddings can preserve geometry types and capture some spatial relations, challenges remain in estimating numeric values

and retrieving spatially related objects due to the loss of magnitude during tokenization [7]. Despite the possibility of ameliorating the issue by modifying notations or applying chain-of-thought prompting [10], this research highlights the need for improvement in terms of capturing the nuances and complexities of the underlying geospatial data and integrating domain knowledge to support various GeoAI applications using LLMs.


References

- 1 Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.
- 2 Tom B. Brown et al. Language models are few-shot learners, 2020. [arXiv:2005.14165](https://arxiv.org/abs/2005.14165).
- 3 Eliseo Clementini and Paolino Di Felice. A model for representing topological relationships between complex geometric features in spatial databases. *Information sciences*, 90(1-4):121–136, 1996.
- 4 Anthony G Cohn and Shyamanta M. Hazarika. Qualitative spatial representation and reasoning: An overview. *Fundamenta informaticae*, 46(1-2):1–29, 2001.
- 5 Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- 6 Max J Egenhofer. Reasoning about binary topological relations. In *Proceedings of the 2nd Symposium on Advances in Spatial Databases: SSD'91 Zurich, Switzerland, August 28–30*, pages 141–160. Springer, 1991.
- 7 Simon Frieder, Luca Pinchetti, Ryan-Rhys Griffiths, Tommaso Salvatori, Thomas Lukasiewicz, Philipp Christian Petersen, Alexis Chevalier, and Julius Berner. Mathematical capabilities of chatgpt. *arXiv preprint arXiv:2301.13867*, 2023.
- 8 Renzhong Guo. Spatial objects and spatial relationships. *Geo-spatial Information Science*, 1(1):38–42, 1998.
- 9 Krzysztof Janowicz, Song Gao, Grant McKenzie, Yingjie Hu, and Budhendra Bhaduri. Geoai: spatially explicit artificial intelligence techniques for geographic knowledge discovery and beyond. *International Journal of Geographical Information Science*, 34(4):625–636, 2020.
- 10 Guillaume Lample and François Charton. Deep learning for symbolic mathematics. *arXiv preprint arXiv:1912.01412*, 2019.
- 11 Zhenlong Li and Huan Ning. Autonomous gis: the next-generation ai-powered gis. *arXiv preprint arXiv:2305.06453*, 2023.
- 12 Gengchen Mai, Weiming Huang, Jin Sun, Suhang Song, Deepak Mishra, Ninghao Liu, Song Gao, Tianming Liu, Gao Cong, Yingjie Hu, et al. On the opportunities and challenges of foundation models for geospatial artificial intelligence. *arXiv preprint arXiv:2304.06798*, 2023.
- 13 Gengchen Mai, Krzysztof Janowicz, Yingjie Hu, Song Gao, Bo Yan, Rui Zhu, Ling Cai, and Ni Lao. A review of location encoding for geoai: methods and applications. *International Journal of Geographical Information Science*, 36(4):639–673, 2022.
- 14 Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.
- 15 David A Randell, Zhan Cui, and Anthony G Cohn. A spatial logic based on regions and connection. *KR*, 92:165–176, 1992.
- 16 Markus Reichstein, Gustau Camps-Valls, Bjorn Stevens, Martin Jung, Joachim Denzler, and Nuno Carvalhais. Deep learning and process understanding for data-driven earth system science. *Nature*, 566(7743):195–204, 2019.
- 17 Maria Vasardani, Stephan Winter, and Kai-Florian Richter. Locating place names from place descriptions. *International Journal of Geographical Information Science*, 27(12):2509–2532, 2013.
- 18 Yang Wang, Huilin Peng, Yiwei Xiong, and Haitao Song. Spatial relationship recognition via heterogeneous representation: A review. *Neurocomputing*, 2023.

Framework for Motorcycle Risk Assessment Using Onboard Panoramic Camera

Natchapon Jongwiriyanurak¹ ✉ 

Department of Civil, Environmental and Geomatic Engineering, University College London, UK

Zichao Zeng ✉ 


Department of Civil, Environmental and Geomatic Engineering, University College London, UK

Meihui Wang ✉ 

Department of Civil, Environmental and Geomatic Engineering, University College London, UK

James Haworth ✉ 

Department of Civil, Environmental and Geomatic Engineering, University College London, UK

Garavig Tanaksaranond ✉ 

Department of Survey Engineering, Faculty of Engineering, Chulalongkorn University, Bangkok, Thailand

Jan Boehm ✉ 

Department of Civil, Environmental and Geomatic Engineering, University College London, UK

Abstract

Traditional safety analysis methods based on historical crash data and simulation models have limitations in capturing real-world driving scenarios. In this experiment, panoramic videos recorded from a motorcyclist's helmet in Bangkok, Thailand, were narrated using an image-to-text model and then put into a Large Language Model (LLM) to identify potential hazards and assess crash risks. The framework can assess static and moving objects with the potential for early warning and incident analysis. However, the limitations of the existing image-to-text model cause its inability to handle panoramic images effectively.

2012 ACM Subject Classification Information systems → Geographic information systems; Computing methodologies → Scene understanding

Keywords and phrases Traffic incident risk, Large Language Model, Vision-Language Model

Digital Object Identifier 10.4230/LIPICs.GIScience.2023.44

Category Short Paper

Funding This research was supported by UCL Global Engagement and Faculty of Engineering, Chulalongkorn University Fund.

1 Background

Traffic incidents are a global issue that causes significant economic and social costs. Every year, millions of people die or are injured in road crashes worldwide, costing countries 3% of their Gross Domestic Product (GDP) on average, with most incidents happening in low- and middle-income countries [13]. According to the World Health Organisation (WHO), Thailand has one of the world's highest road traffic fatality rates, with an average of 22,000 deaths annually. Bangkok, the capital city, is a hotspot for traffic incidents, with almost 1 million casualties a year in 2020 and 2021, 90% of whom were motorcyclists. Tracking the cause of these incidents is challenging as they can often be attributed to multiple factors.

¹ corresponding author



Safety analysis approaches for local roads in Bangkok face limitations due to incompleteness, unavailability, under-reporting, and a lack of comprehensive crash-related factors and behavioural information [3, 11]. Studies utilise advanced cameras, benefiting from improved camera quality, computational power, and AI integration for street scene analysis. However, reliance on static cameras may limit their ability to capture the complexities of real-world scenarios [4, 5]. Recent progress in visual understanding, particularly in Vision-Language (VL) models and Large Language Models (LLMs), has shown great potential in analysing image-text pairs. This opens up opportunities to leverage pre-training VL and LLMs for evaluating real-time videos and assessing the scenes and behaviours of motorcyclists, thereby enhancing traffic risk assessment [8, 2, 12].

This study presents a framework for investigating the hazardous environment and interactions involving motorcycle riders and their surroundings. Initially, panoramic videos will be recorded in Bangkok, Thailand, using a GoPro Max camera mounted on the rider's helmet. Subsequently, Image-to-Text, video captioning, and LLM will be integrated to extract valuable information to identify potential hazards.

This paper is organised as follows. Section 2 lists the related works before Section 3 elaborates on the methodology and experiment. The preliminary results are described in Section 4. The paper finishes with the conclusion and potential applications of this study.

2 Related works

2.1 Traditional Traffic Risk Assessment

Traditionally, safety analysis has relied on historical crash data [11], which unfortunately may suffer from limitations such as incompleteness, unavailability, under-reporting, and a lack of comprehensive behavioural information, as well as the omission of important crash-related factors [3]. Simultaneously, simulation methods may not accurately represent non-lane-based mixed traffic conditions [1]. Although recent advancements in camera quality, Artificial Intelligence (AI), and computational power have enabled the development of simulation models for driving behaviour at intersections, most studies still heavily rely on static cameras, which may not fully capture the intricate complexities of real-world driving scenarios [4, 5].

2.2 Large Vision and Language Pre-trained Models in Traffic Scenes

Since the introduction of Contrastive Language-Image Pre-training (CLIP), VL Pre-training (VLP) models have rapidly advanced, relying on large text-image datasets [6, 9, 7]. Large VLP models achieve competitive performance on benchmark datasets, even without specific training, through zero-shot learning [9, 6]. Additionally, they have the capability for zero-shot Visual Question Answering (VQA) in the context of traffic image understanding [14]. However, while these large VLP models can efficiently extract textual information from image features, they face challenges when it comes to correlating relevant textual information and performing deeper interpretation, particularly in complex scenes involving multiple objects.

After OpenAI proposed ChatGPT [14], the use of LLMs expanded to specific tasks, excelling in summarising prompts and completing questions, explanations, and captions. However, LLMs lack the ability to extract visual features. By combining LLMs with VLP models, they can effectively interpret text features from images and gain detailed information through VQA. This combination overcomes the limitations of large VLP models in explaining phenomena and the inability of LLMs to extract image features without additional training. It is like a visually impaired person relying on an interpreter for specific tasks to understand



■ **Figure 1** Paronomic video dataset coverage in Bangkok, Thailand (left) and an example of object detection (right).

their surroundings. While Large Language-and-Vision Assistant (LLaVA) has shown good interpretation of scenes in many scenarios [7], it falls short in object tracking in videos. To address this, a novel framework proposed in this study incorporates object detection and instance segmentation at each keyframe before applying VLP models. This framework aims to interpret VLP models for analysing traffic scenes, with potential applications in real-time traffic interpretation for early warning of potential risks and incident analysis by stakeholders and planners.

3 Methodology and Experiment

3.1 Data collection

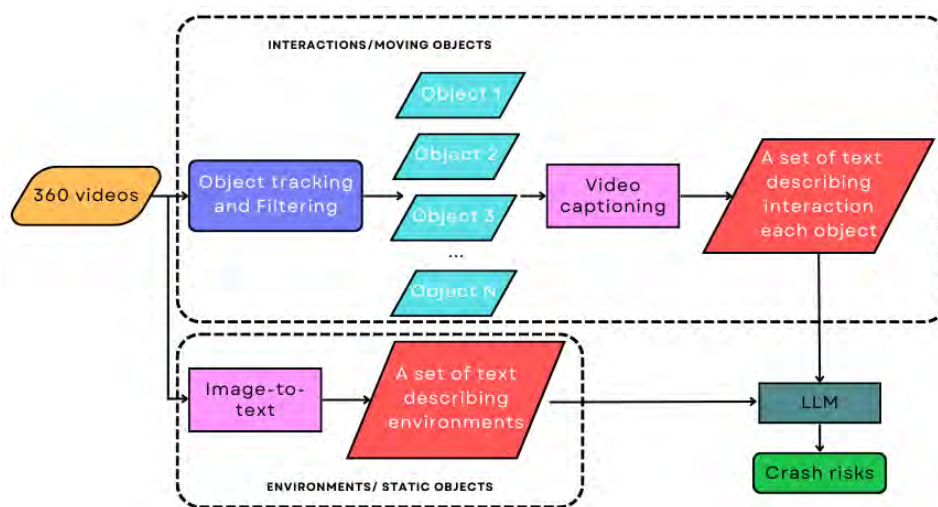
This study involved collecting panoramic videos using a GoPro Max camera mounted on a helmet while riding a motorcycle from December 18, 2022, to January 16, 2023. The camera was set to 360 video mode with 5.6k resolution and 30 Frames per second. The journeys mostly covered the route from home in the Phaya Thai district to Chulalongkorn University in central Bangkok, as indicated by the green dots on the map with examples of captured scenes in Figure 1 (left) with an example of object detection (right). It is important to note that these journeys were routine activities and did not put the user at increased risk. The study obtained ethics approval from UCL and Chulalongkorn University. The dataset is the 360-view of street scenes from a motorcyclist's helmet across Bangkok's streets during diverse times of day, including peak and off-peak periods on weekdays and weekends.

3.2 Framework

This section introduces the proposed framework for identifying motorcycle crash risks from panoramic videos outlined in section 3.1. The framework overview is shown in figure 2 and considers two types of objects: environment or static objects and interactions or moving objects by using different VL models.

The environments or static objects are described using an image-to-text model, which provides information on non-moving objects, such as the number of lanes, flow density, road surface conditions, weather, and lighting.

The framework tracks and filters the interactions between the rider and surrounding objects. It excludes smaller boxes far from the rider to reduce computational costs and contribute less to risk. This exclusion is done to reduce computational costs. Each object is



■ **Figure 2** Framework for panoramic video crash risk analysing.

considered separately within the model, with other objects blurred. This approach focuses on capturing the interaction between the object itself and the rider. A video captioning model is employed to generate descriptions for each moving object. This model processes specific seconds of video footage and generates a set of sentences that describe the interactions between the surrounding objects and the motorcyclist.

The descriptions generated from both moving and static objects are input into a LLM to assess the potential crash risks. The existing LLM will be implemented and fine-tuned to rate a quantitative score that quantifies the level of risk. The visual risk score obtained from the LLM is then combined with trajectory information derived from the camera's GPS, which integrates with Geographic Information System (GIS) data. This GIS data may include historical incident records and Points-of-Interest. Through this integration, the final risk score is computed. This framework can potentially alert the rider when the risk score surpasses a predefined threshold. Such a system could be a valuable tool for motorcyclists, particularly when the camera is equipped with a GPU for real-time processing capabilities.

The ongoing study aims to implement image-to-text, video captioning and LLMs for risk rating purposes. As part of this framework, the image-to-text approach was tested using the LLaVA model [7]. The LLaVA model prompts 6 questions to gather information on various factors that are considered critical in assessing motorcycle crash risks. These factors include flow density, number of lanes, weather conditions, traffic signs (specifically speed limits), road surface conditions, and lighting conditions. By incorporating these crucial risk factors, the study seeks to enhance the accuracy and effectiveness of the risk rating process by validating 16 and 8 images during day and night time, respectively. Researchers manually supervise by rating 1 as correct, 0.5 as partially correct and 0 as wrong, then will calculate the accuracy of returning captions.

4 Preliminary results

The preliminary results are presented in table 1, revealing the accuracy from 6 prompts against manual supervision. The model performed well in classifying flow density, weather, and lighting. However, it showed relatively poor performance in identifying the number of lanes, road surface conditions, and traffic signs. Additionally, the model showed slight differences in performance between day and night time.

■ **Table 1** Accuracy (in %) of captions tested against manual supervision.

Class	% (all)	% (day)	% (night)
Number of lanes	39.6	40.6	37.5
Flow density	81.3	75.0	93.8
Road surface	25.0	37.5	0.0
Traffic sign	16.7	15.6	18.8
Weather	77.1	75.0	81.3
Lighting	83.3	75.0	100.0

Lighting conditions are the easiest to identify, as shown in table 1. This is attributed to the straightforward evaluation of lighting based on red, green, and blue (RGB) values during visual decoding. The model also demonstrated high accuracy in identifying flow density. This is because VLP models have learned vehicles from the abundance of traffic images for a large dataset, and LLM can also easily understand traffic congestion at a textual level. The model effectively determined the weather conditions due to the substantial coverage of sky or weather in traffic images, as the model was pre-trained in a large weather-related sample size.

However, The accuracy of counting the number of lanes is relatively poor in panoramic images, which introduce horizontal misalignment between lane lines and vehicles, leading to confusion for the pre-trained model. Previous work has shown that image understanding tasks trained mainly on rectilinear images benefit from re-projecting equirectangular images to rectilinear before the visual task is performed [10]. Identifying road surface conditions poses a challenge due to the diverse range of colours, conditions, and materials found in different countries. This difficulty is exemplified by the misclassification of cement surfaces as wet surfaces in this study. The most challenging class to identify is traffic signs. It is worth noting that traffic signals (red/green/yellow lights) were considered traffic signs in the textual decoding, and tail lights from vehicles are often labelled as traffic signals, further contributing to the difficulty in accurately identifying traffic signs.

5 Conclusion and Future work

In this study, we proposed a framework to examine motorcycle incident risk by using VLP and LLM models from a panoramic video dataset. The video data was collected in Bangkok, Thailand, by mounting a 360 camera on a motorcyclist's helmet to record the interaction between the surroundings and the rider. A VLP model, LLaVA, is tested on a series of panoramic images in the daytime and nighttime. Promptings related to traffic incident risks are used. The results show the potential of using the pre-trained model to describe safety related features, from testing flow density, weather and lighting conditions, and images for prompting the LLM to rate the incident risk. On the other hand, the results reveal the limitations of using panoramic images when counting the number of lanes, road surfaces, and traffic signs.

In future developments, the framework will incorporate distortion correction to mitigate potential misinterpretations caused by distorted geometries. The objective is to describe critical risks associated with stationary objects and environments accurately. Moreover, there will be a strong emphasis on understanding traffic scenes within the framework model, achieved through the fine-tuning and training of pre-training VL and LLM for visual traffic comprehension and textual analysis. The risk analysis will transition from image-to-text to video captioning, integrating the detection and tracking of moving objects. Unrelated objects

will be disregarded or given lower weights using depth estimation techniques to enhance accuracy. The overarching goal of this comprehensive framework is to comprehend crash risks for motorcyclists and provide real-time notifications to the rider when equipped with graphics processing units on the panoramic camera or edge device. While the framework holds the potential for transferability to other cities, careful consideration must be given to factors such as the environment, vehicles, behaviours, and contextual risks.

References

- 1 Gowri Asaithambi, Venkatesan Kanagaraj, and Tomer Toledo. Driving Behaviors: Models and Challenges for Non-Lane Based Mixed Traffic. *Transportation in Developing Economies*, 2(2):19, October 2016. doi:10.1007/s40890-016-0025-6.
- 2 Jun Chen, Deyao Zhu, Kilichbek Haydarov, Xiang Li, and Mohamed Elhoseiny. Video ChatCaptioner: Towards Enriched Spatiotemporal Descriptions, April 2023. arXiv:2304.04227 [cs]. URL: <http://arxiv.org/abs/2304.04227>.
- 3 Rupam Deb and Alan Wee-chung Liew. Missing Value Imputation for the Analysis of Incomplete Traffic Accident Data. In Xizhao Wang, Witold Pedrycz, Patrick Chan, and Qiang He, editors, *Machine Learning and Cybernetics*, volume 481, pages 275–286. Springer Berlin Heidelberg, Berlin, Heidelberg, 2014. Series Title: Communications in Computer and Information Science. doi:10.1007/978-3-662-45652-1_28.
- 4 Nopadon Kronprasert, Chomphunut Sutheerakul, Thaned Satiennam, and Paramet Luathep. Intersection Safety Assessment Using Video-Based Traffic Conflict Analysis: The Case Study of Thailand. *Sustainability*, 13(22):12722, November 2021. doi:10.3390/su132212722.
- 5 Gabriel Lanzaro, Tarek Sayed, and Rushdi Alsaleh. Can motorcyclist behavior in traffic conflicts be modeled? A deep reinforcement learning approach for motorcycle-pedestrian interactions. *Transportmetrica B: Transport Dynamics*, 10(1):396–420, December 2022. doi:10.1080/21680566.2021.2004954.
- 6 Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models, May 2023. arXiv:2301.12597 [cs]. URL: <http://arxiv.org/abs/2301.12597>.
- 7 Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual Instruction Tuning, April 2023. arXiv:2304.08485 [cs]. URL: <http://arxiv.org/abs/2304.08485>.
- 8 Jesus Perez-Martin, Benjamin Bustos, Silvio Jamil F. Guimarães, Ivan Sipiran, Jorge Pérez, and Grethel Coello Said. A Comprehensive Review of the Video-to-Text Problem, November 2021. arXiv:2103.14785 [cs]. URL: <http://arxiv.org/abs/2103.14785>.
- 9 Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision, February 2021. arXiv:2103.00020 [cs]. URL: <http://arxiv.org/abs/2103.00020>.
- 10 E. Sanchez Castillo, D. Griffiths, and J. Boehm. SEMANTIC SEGMENTATION OF TERRESTRIAL LIDAR DATA USING CO-REGISTERED RGB DATA. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLIII-B2-2021:223–229, June 2021. doi:10.5194/isprs-archives-XLIII-B2-2021-223-2021.
- 11 Chamroeun Se, Thanapong Champahom, Sajjakaj Jomnonkwo, and Vatanavongs Ratanavaraaha. Motorcyclist injury severity analysis: a comparison of Artificial Neural Networks and random parameter model with heterogeneity in means and variances. *International Journal of Injury Control and Safety Promotion*, pages 1–16, June 2022. doi:10.1080/17457300.2022.2081985.
- 12 Junke Wang, Dongdong Chen, Chong Luo, Xiyang Dai, Lu Yuan, Zuxuan Wu, and Yu-Gang Jiang. ChatVideo: A Tracklet-centric Multimodal and Versatile Video Understanding System, April 2023. arXiv:2304.14407 [cs]. URL: <http://arxiv.org/abs/2304.14407>.

- 13 WHO. Global status report on road safety 2018. Technical Report 2, WHO, 2018. ISBN: 9789290496977 ISSN: 00142972 Publication Title: World Health Organization Volume: 3. doi:10.18041/2382-3240/saber.2010v5n1.2536.
- 14 Ou Zheng. ChatGPT Is on the Horizon: Could a Large Language Model Be All We Need for Intelligent Transportation? *Computation and Language*, March 2023. doi:10.48550/arXiv.2303.05382.

National-Scale Spatiotemporal Variation in Driver Navigation Behaviour and Route Choice

Elliot Karikari¹  

Leeds Institute for Data Analytics, University of Leeds, UK

Manon Prédhumeau  

School of Geography, University of Leeds, UK

Peter Baudains  

ESRC Consumer Data Research Centre, University of Leeds, UK

Ed Manley  

School of Geography, University of Leeds, UK

Abstract

Understanding human behaviour is an integral task in GIScience, facilitated by increasingly large and descriptive datasets on human activity. Large-scale trajectory data have been particularly useful in measuring behaviours in different contexts, and understanding the relationship between the built environment and people. Yet, to date, most of these studies have focused on urban or regional scale analyses, with less exploration of behavioural variation at larger spatial scales. Human navigation behaviour is inherently linked to variation in spatial structure, and a study of national variations could help to better understand this variability. In this paper, we analyse GPS data from over 1 million journeys by 50,000 connected cars across the UK. Some key statistics relating to route choice are computed, and their variations are explored over time and space. A k-mean clustering of the trips identifies different types of trips and shows that their distribution varies by time of day and across the country. The insights gained from the data highlight spatio-temporal variations in road navigation, which should be considered in transportation modelling and planning.

2012 ACM Subject Classification Applied computing → Transportation

Keywords and phrases Connected Car, Geospatial big Data, Navigation Behaviour, Cluster Analysis

Digital Object Identifier 10.4230/LIPIcs.GIScience.2023.45

Category Short Paper

Acknowledgements The data for this research have been provided by the Consumer Data Research Centre, an ESRC Data Investment, under project ID CDRC 376, ES/L011840/1; ES/L011891/1.

1 Introduction

The increasing availability of vehicle usage data, made possible by the rise of electric connected vehicles, presents an opportunity for researchers to analyse vast amounts of data related to speed, location, and direction [2]. The ubiquity of the technology means that never before have granular data on navigation behaviour been available at such a large scale. In this study we leverage connected car data to gain novel insights into human mobility patterns and behaviour, scaling analysis up to the national scale. While previous studies have explored various aspects of mobility, such as travel distances, radius of gyration, and visited locations [6], this study specifically examines drivers' routes. Studies have relied on diverse data sources, including GPS tracking devices [6, 5], mobile phone data [4], and transportation surveys [8], revealing insights into the fundamental drivers of navigation behaviour. To date,

¹ corresponding author



there has been no examination of navigation behaviour across an entire country. This study aims to fill this gap by examining the driver navigation behaviour and route choice using national-scale GPS data. The central research question of this study is: to what degree do drivers' navigation behaviours and route choices in the UK vary spatially and temporally?

In this paper, we explore navigation across the entire UK. We observe how a set of indicators describing navigation behaviour vary over space and time. We outline the methods and data involved in the study, before describing the results and their implications.

2 Method

The methods used to derive insights into navigation behaviour are outlined in our previous work [3]. This paper established a methodology for deriving six key statistical measures - travel distance, travel time, stop time, number of turns, angular deviation and sinuosity - with application to the same data. Here, we extend our previous work by applying these six key statistical measures to 1,224,270 trips, i.e. 66.92% of the entire dataset and analysing their spatio-temporal variations. Section 3 details the processing undertaken to clean up the data retaining only one-way trips over 4 weeks in July. Section 4 then presents the statistical results obtained and a k-mean clustering analysis, in order to identify patterns associated with the different journey types and to examine their spatial and temporal variation.

3 Data

This work uses high-frequency GPS recordings from 50,000 connected cars across the UK (<https://data.cdrc.ac.uk/dataset/wejo-connected-vehicle-trajectories>). The dataset consists of over 400 million GPS data points, collected in July 2022 where an observation was recorded every 3 seconds on average during each journey (over 1.8 million). To ensure anonymity, the first and last 15 seconds of each journey have been removed.

An initial two-stage filtering of the data was applied. We selected a four-week period from 4th to 31st July for the analysis. This timeframe provides a balanced selection of weekdays and weekends. Some trips in the dataset were then identified as round trips, which are defined as trips for which the Haversine distance between origin and destination is less than 800 meters. These round trips were filtered out as they were found to greatly skew sinuosity results. Further analysis of this data could reveal specific behaviour associated with round trips. The present analysis however focused on the behaviours associated with one-way journeys which makes up 66.9% of the entire dataset.

In this study, we used Python and the Scikit-mobility library [7] to process data and generate key statistics such as travel distance and stop time. We also derived additional measures such as travel time, number of turns, cumulative angular deviation, and sinuosity, which were essential in providing insights into human mobility patterns.

4 Results

The results computed on the 1,224,270 trips indicate a diverse set of navigation behaviours within the data. Table 1 shows a summary of descriptive statistics generated per journey.

Results reveal a wide range of variability in distances travelled. On average, drivers tended to travel 13.1 km. However, this average is skewed by a few long trips as half of all trips made were below 5.5 km. It was also shown that stop time accounted for approximately 28% of their overall travel time. Furthermore, we observed that people tended to take routes

■ **Table 1** Descriptive statistics on 1,224,270 trips.

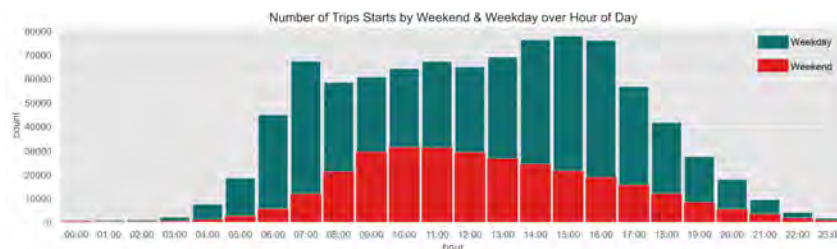
	Min	Max	Mean	Q1	Median	Q3	SD
Travel distance (km)	0.8	715.4	13.1	2.8	5.5	12.6	24.9
Travel time (min)	0.5	747.2	16.7	5.7	10.30	19.5	20.5
Stop time (min)	0.0	646.3	4.7	0.0	2.3	5.7	7.8
Number of turns	0.0	863.0	17.1	8.0	13.0	22.0	15.1
Cumulative angular deviation (°)	0.5	111197.7	2504.3	1189.9	1955.2	3168.1	2074.5
Sinuosity	1.00	273.16	1.59	1.23	1.37	1.60	1.62

with an average of 17 turns per journey. This suggests that the complexity of travel routes should be considered when analysing travel behaviour. Finally, the average sinuosity of 1.59, meaning routes are around 60% longer than the Haversine distance, highlights a considerable amount of inefficiency in navigation behaviour and/or infrastructure. For comparison, [1] reported average sinuosities of 1.377 in Boston and 1.339 in San Francisco for pedestrians.

4.1 Variation over space and time

Next, we accessed spatiotemporal variation, the count of trip starts and sinuosity by time of the day and its geographic variation.

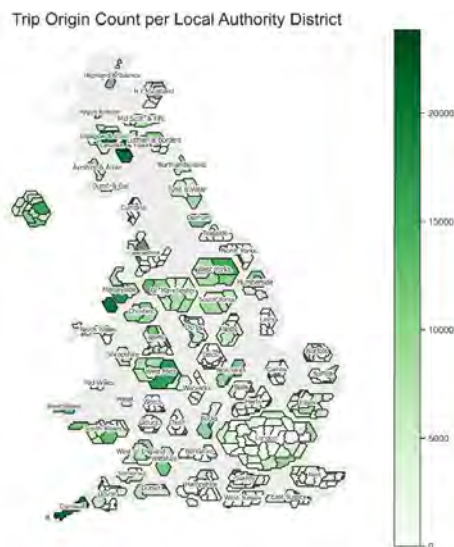
Figure 1 shows high trip starts recorded during all hours of the day on weekdays (in green). Two distinct peaks were identified, at 07:00, and between 14:00 to 16:00. The morning peak is likely indicative of people going to work, while the afternoon peak may be attributed to picking up children from school or people leaving work. The number of trips starts on the weekend (in red), steadily increased during the early hours of the day peaking at 10:00, before steadily declining towards the end of day.



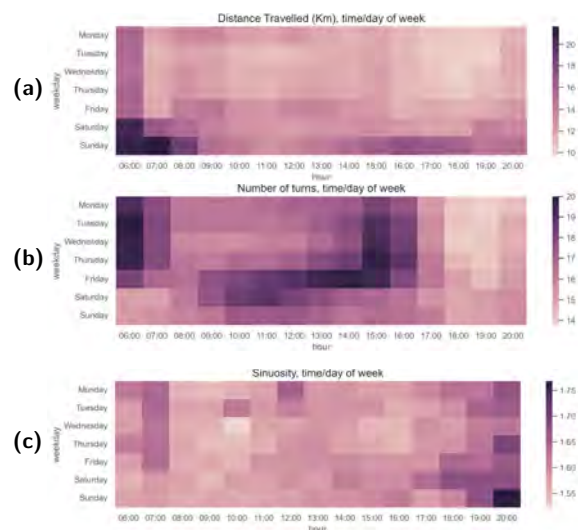
■ **Figure 1** Distribution of connected car trip origins over time.

Geographic visualisations were done using a non-contiguous cartogram at the Local Authority (LA) level from <https://github.com/houseofcommonslibrary>. The LAs have been grouped and scaled in size relative to their populations. Figure 2 shows some regions with a high number of trip origins in Scotland such as Lanarkshire and Falkirk, and Glasgow and Clyde. Cornwall in the Southwest, parts of West Midlands, Merseyside, and West Yorkshire are also highly represented.

The sinuosity variable measures how much a trip deviates from the Haversine distance between the origin and destination. Trips with a sinuosity of 1 are direct and identical to their equivalent Haversine distance. Trips in Bedfordshire, Northamptonshire, Lanarkshire and Falkirk, Wiltshire, and Tyne and Wear have a relatively high average sinuosity (from 2.5 to 2.8). This suggest that drivers in these regions are constraints by infrastructure into driving further to reach their destinations, or that drivers take detours to avoid congestion.



■ **Figure 2** Distribution of connected car trip origins.



■ **Figure 3** Average (a) distance travelled, (b) number of turns and (c) sinuosity over time and day of week.

Within areas with relatively lower sinuosity (from 1.7 to 1.9), i.e. Gloucestershire, Somerset, Highland and Islands, Oxfordshire, and Mid Wales, more direct routes are possible. Camden in the London area has a very high sinuosity (>8.5), which calls for future investigation.

Distances travelled vary depending on the day of the week and the time of day (Figure 3a). Long trips are more common in the early morning hours (around 06:00) on all days of the week, with more long trips on weekends starting between 06:00 and 07:00. This indicates a self-selection bias, in that people who need to travel further are more likely to be leaving in the early morning, relative to later in the day. Results also indicate a relationship between the number of turns per trip and start time (Figure 3b). It appears people may opt to take more turns during peak weekday and weekend periods. However, there is no clear relationship between time of day and sinuosity (Figure 3c), meaning that the routes do not deviate more significantly than usual during these periods. This is an indication that people seek to avoid traffic congestion during peak periods, but do not deviate widely from the shortest route.

4.2 Clustering analysis

After exploring the variability in the travel behaviour data, we identified different route types using k-means clustering analysis. This machine learning method groups similar data points together based on their features. Highly correlated route attributes (correlation coefficient > 0.7 or variance inflation factor > 2.5) were not used. As a result, only three variables – travel distance, number of turns, and sinuosity – were used to cluster the trips. We evaluated different values of k (i.e., [2-8]), using silhouette scores and silhouette visualisers, and found that $k=4$ resulted in the most distinct trip types. However, clustering with $k=6$ produced silhouette scores almost as good as $k=4$, and may be worth further exploration.

Short one-way trips (Cluster 0) Direct trips with the fewest average number of turns (11). Observed travel distance (7 km) is on average 56% longer than the Haversine distance. Most trips fell within this cluster (79%).

Mid-range one-way trips (Cluster 1) Longer trips with more turns (37). Observed travel distance (22 km) is on average 93% longer than the Haversine distance. This cluster accounted for 18.9 % of all trips.

Long one-way trips (Cluster 2) Very long trips (139 km), on average 40% longer than the Haversine distance. The average distance travelled is 6 times longer than in Cluster 1 but has only 13% more turns. This may suggest that Cluster 2 uses more major roads. 2.4% of all trips were accounted for in this cluster.

Sinuous trips (Cluster 3) 0.1% of clustered trips were identified as round trips, with unusually high sinuosity values (32) compared to the other clusters (1 to 2). This indicates that the simple filtering process used (removing trips with origin-destination distance <800m) could be improved using a filter to remove high sinuosity trips.

As most trips (79%) were short one-way trips (Cluster 0), we ran the clustering on this subset to identify variations within trips (Table 2).

■ **Table 2** Descriptive statistics of re-clustered Cluster 0.

	Average travelled distance	Average number of turns	Average sinuosity	% trips
0A	6.90	17.66	1.45	34.7
0B	4.06	6.81	1.34	51.4
0C	5.22	13.26	2.66	6.6
0D	29.49	12.99	1.33	7.3

Cluster 0A Short sinuous one-way trips, with high number of turns. Observed travel distance is 45% longer than its Haversine distance.

Cluster 0B Short, direct, low sinuosity, one-way trips. Observed travel distance is 34% longer than its Haversine distance.

Cluster 0C Short highly sinuous one-way trips, with moderate number of turns. Observed travel distances is 166% longer than its Haversine distance.

Cluster 0D Mid-range, low sinuosity, one-way trips with moderate number of turns. Observed travel distance is 33% longer than its Haversine distance.

Further analysis found that Fridays had the highest number of short sinuous trips (0A), while Wednesdays had more short direct trips (0B). Drivers may be more willing to take indirect routes on Fridays when they have more time or are less constrained by work schedules. The analysis also showed that short sinuous trips (0A) were more common in major urban conurbations including London, Greater Manchester, and South Yorkshire (Figure 4), while short direct trips (0B) were highly represented in all other areas. Travel behaviour in some areas may be different from others, possibly due to road infrastructure or specific traffic conditions. Further research could explore these variations and identify potential solutions for improving travel efficiency.

5 Conclusion

This study provides insights into road navigation behaviour across the UK, based on connected cars data. The analysis of travel distance, number of turns, and sinuosity revealed patterns that vary by time of day and day of the week. The identification of different trip types further highlights the variability in navigation behaviour across the UK. This new perspective on navigation behaviour can supplement the outputs of classic surveys and will be used to create synthetic trip datasets that are representative of observed behaviours.



■ **Figure 4** Most frequent cluster for each local authority: short sinuous one-way trips (0A) in green and short direct one-way trips (0B) in yellow.

Results from this study can inform transportation planning and policy. For example, the finding that 25% of the analysed car trips are shorter than 2.8 km can guide the development of zero-emission local policies by identifying where and when drivers make short trips. Moreover, insights from this study may be used to refine transport models with new behavioural patterns, and help to predict drivers' behaviour. However, the lack of socio-demographic data prevents the assessment of the representativeness of the data. This study is purely observational and further exploration of causation is required. Stop time and point of interest data could enable the investigation of the trip purposes. Overall, this study underscores the potential of using vehicle trajectory data to understand travel decisions.

References

- 1 C Bongiorno, Y Zhou, M Kryven, D Theurel, A Rizzo, P Santi, J Tenenbaum, and C Ratti. Vector-based pedestrian navigation in cities. *Nat. Comput. Sci.*, 1(10):678–685, 2021.
- 2 R Coppola and M Morisio. Connected car: technologies, issues, future trends. *ACM Comput. Surv.*, 49(3):1–36, 2016.
- 3 E Karikari, M Prédhumeau, P Baudains, and E Manley. Analysing connected car data to understand vehicular route choice. In *31st Annual Geographical Information Science Research UK Conference (GISRUK), Glasgow, Scotland, 2023*.
- 4 L Li, S Wang, and F-Y Wang. An analysis of taxi driver's route choice behavior using the trace records. *IEEE Trans. Comput. Soc. Syst.*, 5(2):576–582, 2018.
- 5 EJ Manley, JD Addison, and T Cheng. Shortest path or anchor-based route choice: a large-scale empirical analysis of minicab routing in London. *J. Transp. Geogr.*, 43:123–139, 2015.
- 6 L Pappalardo, S Rinzivillo, Z Qu, D Pedreschi, and F Giannotti. Understanding the patterns of car travel. *Eur. Phys. J. Spec. Top.*, 215:61–73, 2013.
- 7 L Pappalardo, F Simini, G Barlacchi, and R Pellungrini. scikit-mobility: A Python library for the analysis, generation and risk assessment of mobility data. *J. Stat. Softw.*, 2022.
- 8 B Yin and F Leurent. What are the multimodal patterns of individual mobility at the day level in the Paris region? A two-stage data-driven approach based on the 2018 Household Travel Survey. *Transportation*, pages 1–30, 2022.

Status Poles and Status Zoning to Model Residential Land Prices: Status-Quality Trade off Theory

Thuy Phuong Le¹ ✉ 

VNU University of Science, Hanoi, Vietnam

Alexis Comber ✉ 

School of Geography, University of Leeds, UK

Binh Quoc Tran ✉ 

VNU University of Science, Hanoi, Vietnam

Phe Huu Hoang ✉

R & D Consultants, Hanoi, Vietnam

Huy Quang Man ✉

VNU University of Science, Hanoi, Vietnam

Linh Xuan Nguyen ✉

VNU University of Science, Hanoi, Vietnam

Tuan Le Pham ✉

VNU University of Science, Hanoi, Vietnam

Tu Ngoc Bui ✉

VNU University of Science, Hanoi, Vietnam

Abstract

This study describes an approach for augmenting urban residential preference and hedonic house price models by incorporating Status-Quality Trade Off theory (SQTO). SQTO seeks explain the dynamic of urban structure using a multipolar, in which the location and strength of poles is driven by notions of residential status and dwelling quality. This paper presents in outline an approach for identifying status poles and for quantifying their effect on land and residential property prices. The results show how the incorporation of SQTO results in an enhanced understanding of variations in land / property process with increased spatial nuance. A number of future research areas are identified related to the status pole weights and the development of status pole index.

2012 ACM Subject Classification Information systems → Spatial-temporal systems; Applied computing → Economics

Keywords and phrases spatial theory, house prices

Digital Object Identifier 10.4230/LIPICs.GIScience.2023.46

Category Short Paper

1 Introduction

The importance of location in land valuation has been confirmed in many studies[9]. The increasing use of explicitly spatial methods in land valuation is an emergent trend[6]. In urban areas, land value is closely related to spatial structure, such as proximity to central business districts (CBDs)[10]. However, it is difficult to quantify the spatial variation of drivers[6]. Status – Quality Trade Off theory (SQTO) explains the dynamic structure of

¹ Corresponding author



© Thuy Phuong Le, Alexis Comber, Binh Quoc Tran, Phe Huu Hoang, Huy Quang Man, Linh Xuan Nguyen, Tuan Le Pham, and Tu Ngoc Bui;
licensed under Creative Commons License CC-BY 4.0

12th International Conference on Geographic Information Science (GIScience 2023).

Editors: Roger Beecham, Jed A. Long, Dianna Smith, Qunshan Zhao, and Sarah Wise; Article No. 46; pp. 46:1–46:6

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

residential areas using a multipolar model of two components: housing status and dwelling quality[4]. Each pole or center represents the highest degree of attractiveness about a certain type of social status. Housing status is the value of the non-physical (or intangible) factors, including cultural, economic, environmental political, etc., which distinguish different levels of housing desirability. Dwelling quality relates to the physical, measurable elements relating to the normal use of a dwelling[4]. The benefits of applying SQTO have important implications for housing and real estate policies including:

- Refinement of statistical methods and models for analysing the housing market and value forecasting, by including housing status and dwelling quality[1, 2].
- The identification of housing status pole locations, capturing frequently intangible qualities that are inherently associated with the evolving spatial structure of cities.
- The opportunity to capture, explanation and predict future housing bubbles.

Most studies, including in Vietnam, focus on the first of these[7]. This short research paper examines the second and third in land valuation.

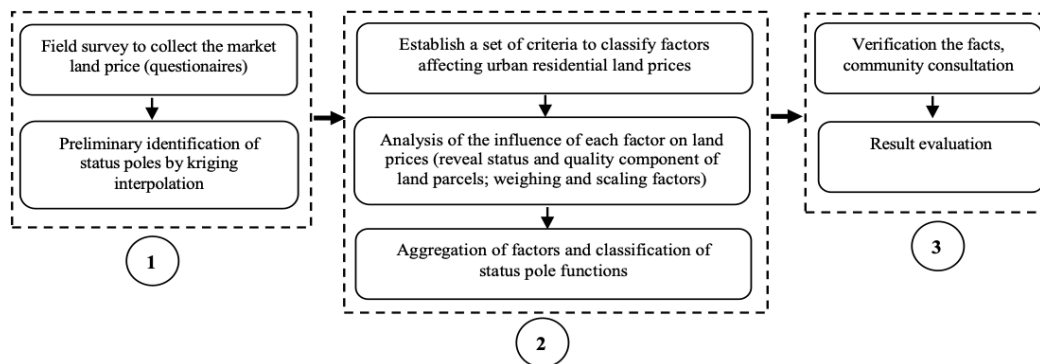
2 Background

SQTO defines a status pole as “the highest point of certain kinds of social status, recognized by a given proportion of the population”[4]. They capture qualitative neighbourhood perceptions such as wealth, political power, business, culture, ethnicity, education, etc, and play an important role in land valuation. Urban residential areas have distinct morphological patterns around status poles[4] and are geographically stratified, providing a potential basis for analysis and modelling. Properties can be grouped into homogenous areas based on factors such as use, physical characteristics with different status poles pulling value in positive or negative directions. For example, areas around the CBDs typically have a higher house price [2, 5] and negative pulls have found around landfill [3]. There are enhanced opportunities to support development policy, planning and real estate regulation by better understanding the location and nature of different status poles and importantly, their effect on value and price. In Vietnam, land acquired for development is subject to a state determined compensatory value. In this, land use change is related to a change in value. A further regulatory aspect is that “rumors” can form virtual status poles, leading to real estate bubbles. In order to explain the mechanism of the bubble, a number of recent studies started to look more closely at the components of land value over temporal and geographic dimensions [6].

This paper identifies the areas around different status poles, as the basis for understanding variations in land value. The status pole is the location with the highest point. “status value” that pulls value in the surrounding area (in a positive or negative direction). This can cause land prices in the surrounding area to increase or decrease. One aspect of the status poles are their ability to represent aspirations of different social groups when they choose their residential location. This suggests the need for different factors to be weighted relative to location in any spatial analysis to identify status poles. Here a classic multi-criteria analysis is used to synthesize, evaluate and understand the relative strengths of emergent status poles.

3 Methodology

To identify status poles revealed preferences and stated preferences are combined. Revealed preference methods involve the quantification of people’s preference through market land value (objective). Stated preference are captured through a set of questions with varying degrees of strength (subjective). The combination of the two approaches is the basis for identifying status poles (Figure 1) and the full method is in Le et al. [8]



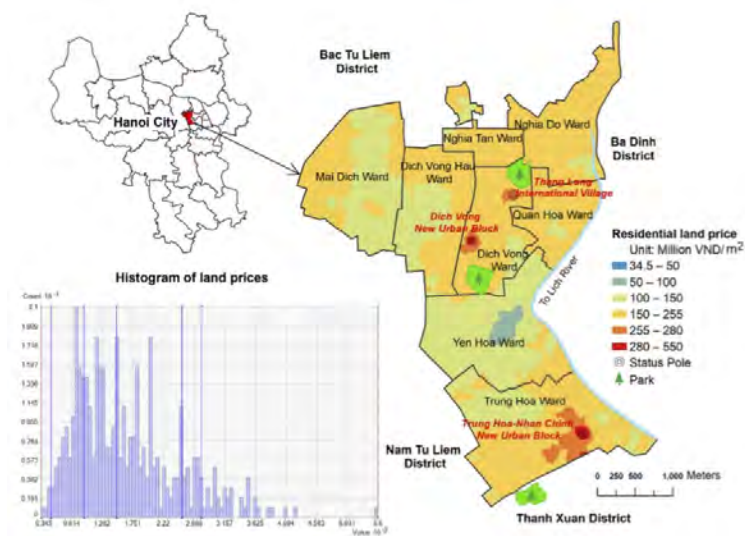
■ **Figure 1** The process of identifying status poles through 3 stages.

The first stage is to delineate areas based areas with the highest and lowest land price, as potential status poles. Questionnaires were used to capture information about residential land parcels sales. These had sections with a total of 46 questions: land owner information (occupation, number of family members, incomes, etc.); land parcel and transaction information (location, area, shape, transfer price, date of transaction, etc.); house information (house type, number of floors, house price, etc.); neighborhood characteristics (water and electricity utilities, security environment, accessibility, etc.). Sample data of land transactions was collected in surrounding areas under normal trading conditions. The minimum required number of samples (N) was estimated as $N > 50 + 8m$ where m is the number of predictors. The samples were interpolated using Kriging to give a spatial distribution of land prices.

The second stage determines the spatial location and function of status poles from analysis of the influence of factors on land prices. These change over time, space and with people's perspective. Criteria were proposed to select locally appropriate factors. In overview, criteria were established for classifying factors affecting urban residential land prices were based on urban quality of life approach with six dimensions (including environmental, physical, mobility, social, psychological, and economic dimensions). A key task of this stage was to determine the weights and scales of influence on land prices of each factor. A variety of methods were explored including network analysis, space syntax, Analytic Network Process, Fuzzy logic, and, for each dimension a composite index (the urban quality of life index - UQoL) is calculated for each land parcel. This provides the basis for determining the function of status poles that were preliminary identified in the first stage.

The third stage is to capture the opinions of people living around the status pole as a form of to verification through questionnaires. A total of 15 questions related to the indicators of urban quality of life, and the attractiveness of status poles were scored on a Likert scale with 5 levels from very dissatisfied to very satisfied. Thus, the three criteria for identifying the status poles are addressed in the three stages of the process.

The case study area is Cau Giay District. This is among the most well-developed districts in Hanoi. It has eight wards and is bordered by old inner districts and new districts. Cadastral maps were collocated with 427 standardized survey samples of real estate transactions from 2017 to 2019. Attribute data was obtained from both field surveys and spatial analyses resulting in each land parcel having 36 attribute fields (such as land market price, shape, frontage, relative position of a parcel to streets, distance from a land parcel to the closest hospital, school, police station, etc.).



■ Figure 2 Preliminary identification of status poles.

4 Results and Discussion

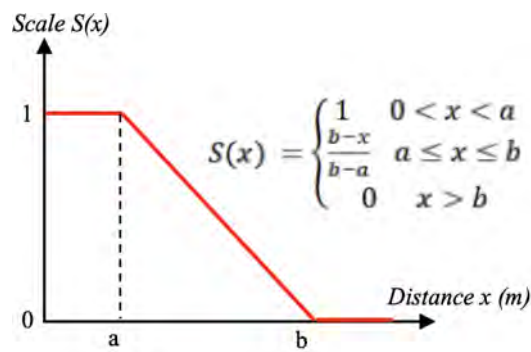
4.1 Identification of status poles

The kriging interpolation generated the spatial distribution of residential land prices, as the basis for preliminary identification of status poles (Figure 2). The results show that residential land prices vary from 34.5 to 550 million VND/m^2 . Three areas with the highest land prices are considered as positive status poles, namely: (1) Thang Long International Village area, (2) Dich Vong New Urban Block (in Dich Vong Ward) and (3) Trung Hoa – Nhan Chinh New Urban Block (in Trung Hoa Ward). The area with the lowest land price is the residential area in Yen Hoa ward, which has a negative status pole.

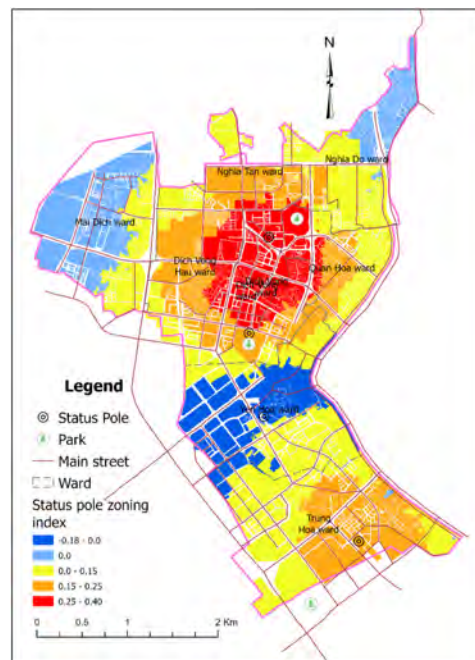
A quality of life index was calculated for each residential land parcel as the basis for determining the function of the status poles. The results show that the area around Cau Giay Park (including Dich Vong New Urban Block) and Thang Long International Village have a high UQoL (0.950-0.995). The area around Trung Hoa – Nhan Chinh New Urban Block with an UQoL index (0.900-0.950) are areas near the park with convenient access to socio-economic locations such as schools, hospitals, offices, etc. Local interviews revealed that the attractiveness of the status poles was 86% in Cau Giay Park, 87.5% in Thang Long International Village, and 76% in Trung Hoa – Nhan Chinh Urban Block. In contrast, the residential area in Yen Hoa ward has a lower UQoL index (0.800) due to poor infrastructure, degraded roads and some locations prone to flooding and cemeteries in this residential area.

4.2 Status pole zones

The interaction between the status poles allows status pole areas to be delineated. According to SQTO are continuous and overlapping rings. In Cau Giay District, there are 3 positive status poles (X, Y, Z) and 1 negative status pole (P). The distance of each land parcel to each status poles was determined (x, y, z, p). Fuzzy logic was used to scale the distance values under the principle that the closer to the status poles, the higher scale (S_x, S_y, S_z, S_p). Figure 3 shows the membership function used for scaling, with the value of a and b depending on individual preferences or as derived from Government regulations. Here these were set as follows: $a = 300m$, $b = 2000m$.



■ **Figure 3** Membership function for scaling distance to status poles.



■ **Figure 4** Result of status pole index zoning in Cau Giay District.

The next step is to calculate status index as follows:

$$I_{status} = (Sx + Sy + Sz + Sp)/4$$

where S_x , S_y , S_z have positive values (+), and S_p has negative value (-). The value of the status index ranges from -1 to 1 to represent positive or negative status poles. The value 0 represents regions not affected by status poles. Figure 4 shows the result of status areas. It can be seen that the linking area between Thang Long International Village and Cau Giay Park (the red area) has the strongest influence in the positive direction. Because these two status poles are located relatively close to each other, they are considered to form a “dual” status pole, with a stronger influence. The blue color represents areas affected by the negative status pole of Yen Hoa Ward. Some areas with light blue color are not affected by all four status poles in Cau Giay District such as Mai Dich Ward, the north of Nghia Do Ward. However, these areas may be partially affected by other status poles in the neighboring areas.

5 Conclusions

This research provides an outline of an approach for identifying status poles related to urban residential land and their effect on price. These were identified as the locations where the influence of qualitative factors on the surrounding area are strongest, causing land prices to increase or decrease sharply and recognized by a given proportion of the population. These influences can be economic, political, environmental, etc., that affect the urban quality of life and their interaction with the status poles form the rings of status pole zones. Future research will to consider the weights of status pole and the application of status pole zoning index in land value. This paper also demonstrates how the concept of status poles SQTO, and the spatial in neighbourhood variation that it captures, can be used to underpin spatially non-stationary house price models. These quantify how the relationship between land value and house price with different factors related to neighbourhood perceptions and the property vary in different parts of the city. Being able to model how and where the processes vary spatially, supports a deeper, more spatially nuanced understanding of the impacts of developments and urban transformation. This is important in locations that are experiencing very rapid urban changes, to identify house price bubbles early, to ensure developments are socially mixed and critically to avoid the commodification of property. The emergence and presence of these of price bubbles can be identified and explained using the proposed Index.

References

- 1 TQ Bui, HN Do, and PH Hoang. House price estimation in hanoi using artificial neural network and support vector machine: in considering effects of status and house quality [paper presentation]. In *FIG Working Week*, 2017.
- 2 A Comber, P Harris, N Quan, K Chi, T Hung, and HH Phe. Local variation in hedonic house price, hanoi: a spatial analysis of sqto theory. In *International Conference on GIScience: Short paper proceedings*, volume 1, pages 54–59, 2016.
- 3 Diane Hite, Wen Chern, Fred Hitzhusen, and Alan Randall. Property-value impacts of an environmental disamenity: the case of landfills. *The Journal of Real Estate Finance and Economics*, 22:185–202, 2001.
- 4 Hoang Huu Phe and Patrick Wakely. Status, quality and the other trade-off: Towards a new theory of urban residential location. *Urban studies*, 37(1):7–35, 2000.
- 5 Hironori Kato and Le Hong Nguyen. Land policy and property price in hanoi, vietnam. *Journal of the Eastern Asia Society for Transportation Studies*, 8:1011–1026, 2010.
- 6 Andy L Krause and Christopher Bitter. Spatial econometrics, land values and sustainability: Trends in real estate valuation research. *Cities*, 29:S19–S25, 2012.
- 7 Thuy P Le. *Research on urban residential land valuation from viewpoint of Status – Quality Trade Off theory, case study in Cau Giay District, Hanoi City*. PhD thesis, VNU University of Science, Vietnam National University, Hanoi Vietnam, 2021.
- 8 Thuy P Le, Phe H Hoang, Linh X Nguyen, Tu N Bui, Tuan L Pham, and Binh Q Tran. Urban quality of life evaluation using land price with status-quality trade-off theory and ecosystem services. *International Journal of Strategic Property Management*, 27(2):92–104, 2023.
- 9 Andreas Ortner, Matthias Soot, and Alexandra Weitkamp. Determining land values by location: Supporting public valuation expert committees in the provision of market transparency. *The role of public sector in local economic and territorial development: Innovation in central, Eastern and South Eastern Europe*, pages 83–96, 2019.
- 10 Paul F Wendt. Theory of urban land values. *Land economics*, 33(3):228–240, 1957.

Investigating MAUP Effects on Census Data Using Approximately Equal-Population Aggregations

Yue Lin  

Department of Geography, The Ohio State University, Columbus, OH, USA

Ningchuan Xiao  

Department of Geography, The Ohio State University, Columbus, OH, USA

Abstract

The modifiable areal unit problem (MAUP) can significantly impact the use of census data as different choices in aggregating geographic zones can lead to varying outcomes. Previous research studied the effects using random aggregations, which, however, may lead to the use of impractical and unrealistic zones that deviate from recommended census geography criteria (e.g., equal population). To address this issue, this study proposes the use of approximately equal-population aggregations (AEPAs) for exploring MAUP effects on various statistical properties of census data, including Moran coefficients, correlation coefficients, and regression statistics. A multistart and recombination algorithm (MSRA) is used to generate multiple sets of high-quality AEPAs for testing MAUP effects. The results of our computational experiments highlight the need for more well-defined census geographies and realistic alternative zones to fully understand MAUP effects on census data.

2012 ACM Subject Classification Computing methodologies → Modeling and simulation

Keywords and phrases Census, heuristics, modifiable areal unit problem, spatial aggregation, spatial autocorrelation

Digital Object Identifier 10.4230/LIPIcs.GIScience.2023.47

Category Short Paper

1 Introduction

The U.S. Census Bureau reports data using a nested hierarchy of geographic zones, beginning with census blocks and progressing to block groups, tracts, counties, and states. The boundaries of these zones are typically created before the digital computer era and are often arbitrary [4], resulting in significant variations in size, population, and demographic makeup [8]. When data is aggregated into different geographic zones and at different scales, statistical properties of the data, such as spatial autocorrelation and correlation coefficients, often demonstrate significant differences from the officially defined zones. This problem is referred to as the modifiable areal unit problem, or MAUP [6], which often causes uncertain and potentially biased census data [5].

A spatial aggregation is a particular way of grouping low-level units (e.g., census blocks) into contiguous high-level units (e.g., census block groups or tracts). Figure 1 illustrates such an aggregation where census block groups (light grey lines) are aggregated into tracts (dark grey lines). During this process, the aggregated data is expected to have different, often reduced, spatial autocorrelation compared to the original data. When using the aggregated data in correlation and regression analysis, the coefficients may also differ from those obtained using the original data.

To understand MAUP effects on the statistical properties of spatial data, algorithms have been developed to generate alternative random aggregations [5, 2]. However, these algorithms often cannot yield spatial aggregations that satisfy some prerequisites of census geography. For example, census tracts in the United States are designed to have a relatively uniform



© Yue Lin and Ningchuan Xiao;
licensed under Creative Commons License CC-BY 4.0

12th International Conference on Geographic Information Science (GIScience 2023).

Editors: Roger Beecham, Jed A. Long, Dianna Smith, Qunshan Zhao, and Sarah Wise; Article No. 47; pp. 47:1–47:6

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany



■ **Figure 1** The official 2010 census block groups (light grey) and tracts (dark grey).

population size of around 4,000 people [9], a criterion that random aggregations typically fail to meet. Using random aggregations, therefore, may only partially reveal MAUP effects on the statistical properties of spatial data.

In the meantime, past research has also advocated the use of geographic zones that are standardized in terms of size, population, and other socioeconomic components [4, 8]. To achieve this, automated zone design algorithms have been developed to generate alternative aggregations that meet these criteria [4, 1, 7]. However, these algorithms often focus on producing one of the many possible aggregations that align with these criteria and may not allow us to fully understand MAUP effects on census data.

This paper reports our work in progress where we utilize a heuristic search method called multistart and recombination algorithm (MSRA) [10] to generate multiple approximately equal-population aggregations (AEPAs). We compare the analysis of MAUP effects derived from AEPAs with that from random aggregations. Specifically, we investigate whether AEPAs significantly differ from random aggregations in (1) spatial autocorrelation, (2) correlation coefficients, and (3) other regression statistics. In the following sections, we first describe the MSRA and then illustrate how AEPAs generated by the MSRA can be used to explore the impact of spatial aggregation on the statistical properties of census data.

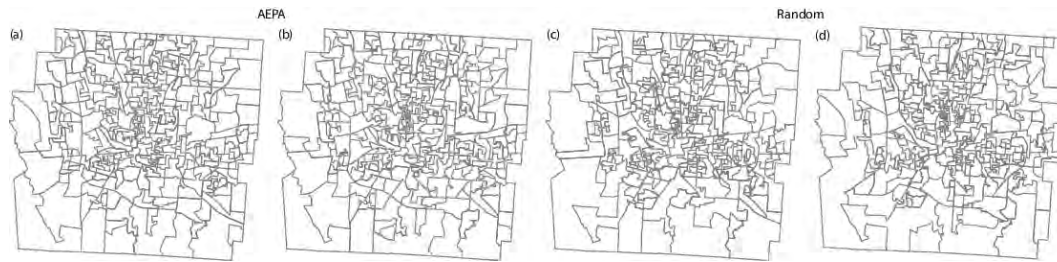
2 The Multistart and Recombination Algorithm

The MSRA is a heuristic search method that identifies a diverse set of high-quality spatial aggregations where the populations in the zones are approximately equal [10]. The algorithm consists of two phases. In the first phase, a multistart process generates a pool of independent aggregations. Each aggregation is randomly created and then improved using an efficient method called the give-and-take algorithm to reduce population differences between zones by swapping units [3]. The second phase is an iterative process where in each iteration two aggregations from the pool are randomly selected and then combined to create a new aggregation. If the new aggregation is new and superior to the worst in the pool, it is added to the pool by replacing the worst aggregation.

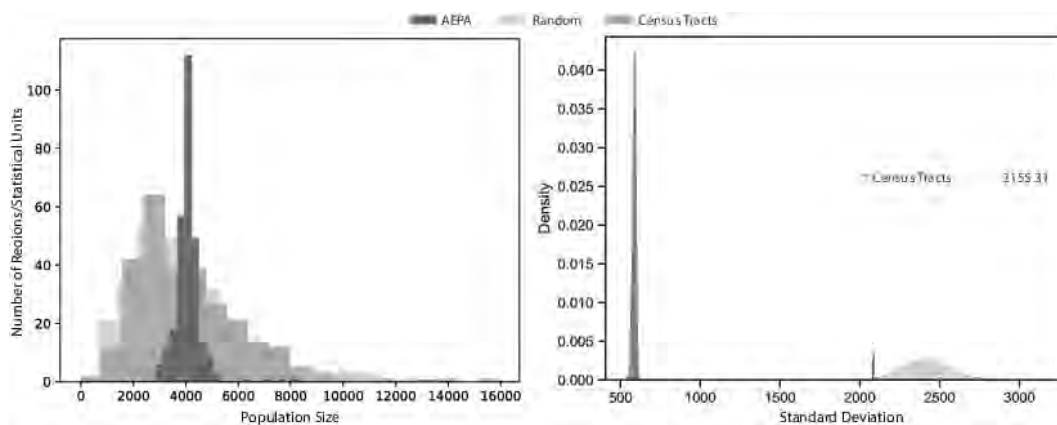
3 Computational Experiments: Design and Results

We chose Franklin County, Ohio as our study area due to its diverse social, economic, and demographic characteristics, as well as its mix of densely populated urban areas and extensive rural areas. The county is composed of 887 block groups that are combined to form 284 tracts (Figure 1). To explore potential alternatives to these census tracts, we use the MSRA to generate 500 AEPAs, each combining 887 block groups into 284 zones with approximately equal population (Figure 2a–b). We also employ the algorithm proposed in [7]

to generate 10,000 random aggregations for comparison (Figure 2c–d). Figure 3 illustrates the population distribution among official census tracts, as well as zones in AEPAs and random aggregations. The results suggest that random aggregations have the greatest variation in population distribution, followed by the official census tracts, and then the AEPAs generated by the MSRA. Using the AEPAs, we investigate MAUP effects on three variables related to the Franklin population: the number of people who work from home (x_1), the number of non-Hispanics (x_2), and the number of people with a Bachelor's degree or higher (y).



■ **Figure 2** Two AEPAs (a, b), and two random aggregations (c, d).



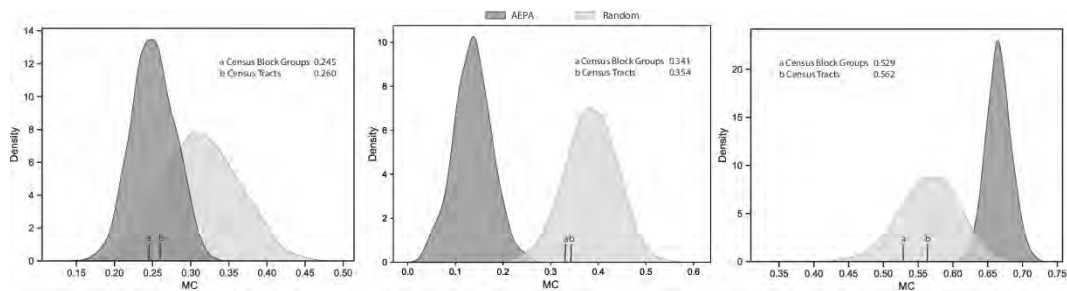
■ **Figure 3** Population distribution and comparison. The left panel displays the population distribution for official census tracts, zones in an AEPA, and zones in a random aggregation. The right panel shows the standard deviation of population size for zones in all AEPAs and random aggregations.

3.1 MAUP effects on Moran coefficients

The Moran coefficient (MC) is a statistical measure that determines the degree of spatial autocorrelation of a variable. Figure 4 demonstrates MAUP effects on the MC using AEPAs and random aggregations, revealing three significant findings. First, the MC resulting from AEPAs can be quite different from the MC at the block group level. For instance, for variable x_2 , the MC at the block group level (0.341) suggests moderate spatial autocorrelation, whereas the average MC resulting from AEPAs (0.136) indicates weak spatial autocorrelation. This finding implies that aggregation can affect the statistical properties of census data. Second, the MC obtained by aggregating data using the official census tracts may not always provide a reliable representation of the MCs that can be obtained through AEPAs. For example, the average MC for variable x_2 obtained through AEPAs is 0.136, while the tract-level MC

47:4 Investigating MAUP Effects on Census Data

is noticeably higher at 0.354. This may seem surprising, but can be explained by the fact that although equal population served as a general principle when the Census Bureau first designed the boundaries of census tracts, these boundaries have not been updated in decades while the population within them has changed substantially. As a result, the population distribution in existing census tracts deviates from equal population, and the statistical properties of census data can also differ from those that can be obtained through AEPAs. Third, it is observed that the distributions of MC under equal population and random aggregations may differ significantly. For example, there is minimal overlap between the distribution of MC under random and equal population aggregations for variables x_2 and y .



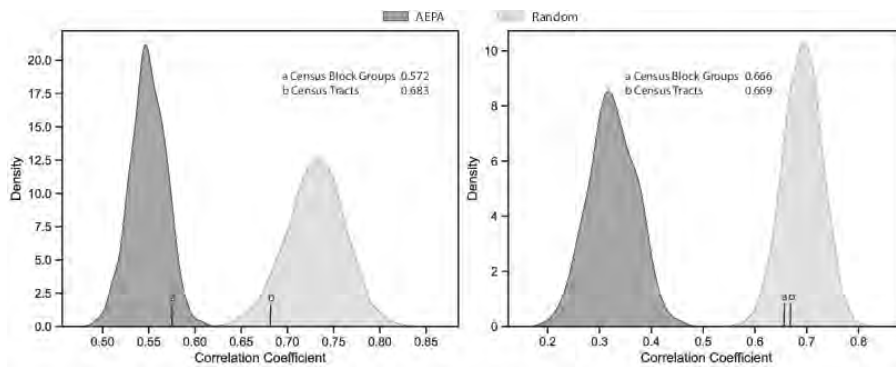
■ **Figure 4** MAUP effects on the MC of variables x_1 (work from home, left), x_2 (non-Hispanics, middle), and y (Bachelor's degree or higher, right).

3.2 MAUP effects on correlation coefficients

Correlation coefficients play a critical role in statistical analysis by quantifying the degree and direction of the relationship between two variables. Figure 5 presents MAUP effects on correlation coefficients between variables x_1 and y and between x_2 and y . The findings are similar to those observed for the MC. First, the correlation coefficients derived using AEPAs can differ substantially from the block group-level coefficient, as demonstrated by the correlation coefficient between x_2 and y being 0.666 at the block group level, whereas the average resulting from AEPAs is substantially lower at 0.325. Second, the correlation coefficients obtained through AEPAs can differ considerably from the tract-level correlation coefficient. For instance, the average correlation coefficient between x_2 and y for AEPAs is 0.325, suggesting a slightly weak association, while the correlation coefficient at the tract level is 0.669, indicating a strong association. Third, the correlation coefficients resulting from AEPAs and random aggregations have little overlap, as previously observed for the MC. This finding reinforces the idea that random aggregations may not yield representative results that reflect MAUP effects when the equal population criterion is applied to modify census geography.

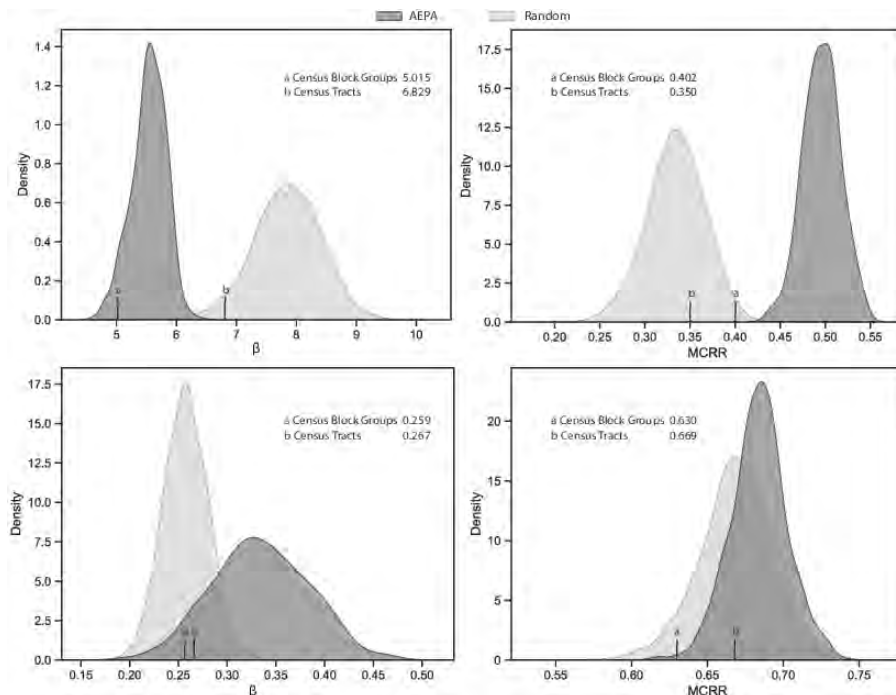
3.3 MAUP effects on regression statistics

Regression analysis is a widely used statistical tool to explore the relationship between a dependent variable and independent variables. In the case of two variables, it is important to consider MAUP effects on the regression slope, which indicates the direction and strength of the association. In addition, if the regression residuals exhibit spatial autocorrelation, the assumption of independent residuals is violated, compromising the validity of linear regression analysis. It is therefore crucial to investigate the spatial autocorrelation of the residuals under different aggregations to determine if they exacerbate or mitigate the issue.



■ **Figure 5** MAUP effects on the correlation coefficients between x_1 (work from home) and y (Bachelor's degree or higher) on the left, and between x_2 (non-Hispanics) and y on the right.

Here, we examine two regression models of the form $y = \alpha + \beta x_1$ and $y = \alpha + \beta x_2$, where α represents the regression intercept and β the regression slope. Figure 6 presents MAUP effects on the regression slope β and the Moran coefficient of the regression residuals (MCRR). It is observed that both β and MCRR exhibit differences using AEPAs compared to block group-level regression statistics. In addition, there are differences between β and MCRR obtained through AEPAs and existing census tracts. Consistent with our findings for MC and correlation coefficient, distributions of β and MCRR generated using AEPAs and random aggregations can have minimal overlap.



■ **Figure 6** MAUP effects on the regression slope β and the MCRR for two models: $y = \alpha + \beta x_1$ (top) and $y = \alpha + \beta x_2$ (bottom).

4 Conclusions

We present a renewed exploration of MAUP effects on univariate and bivariate statistics of census data using multiple approximately equal-population aggregations (AEPAs). Our study yields three key findings. The first highlights the significance of MAUP effects when aggregating census data from low-level units, which can greatly impact the interpretation of statistical properties such as Moran coefficients, correlation coefficients, and regression statistics. Second, the current census geography deviates from the principles that guided its design decades ago, which poses a challenge for understanding and addressing MAUP effects. Our analyses show how existing census tracts, established since 1790 and evolved over time, barely adhere to the equal population criterion today, resulting in statistical properties that differ from what we would expect under equal population aggregation. This finding underscores the need to re-examine the existing census geography and to develop more well-defined geographic zones to help better understand MAUP effects. Finally, our analyses reveal that random aggregations and AEPAs can yield vastly different results regarding MAUP effects. While it is recognized that not all randomly generated sets of zones are suitable for use as census geography, they are still commonly used to study MAUP effects. Our study demonstrates the limitations of such an approach and emphasizes the importance of employing realistic zones with approximately equal population to better understand MAUP effects on census data. Future research can be directed to generalize these findings to other variables and explore the impact of AEPAs in multivariate situations.

References

- 1 Samantha Cockings, Andrew Harfoot, David Martin, and Duncan Hornby. Maintaining existing zoning systems using automated zone-design techniques: methods for creating the 2011 census output geographies for England and Wales. *Environment and Planning A*, 43(10):2399–2418, 2011.
- 2 A Stewart Fotheringham and David WS Wong. The modifiable areal unit problem in multivariate statistical analysis. *Environment and Planning A*, 23(7):1025–1044, 1991.
- 3 Myung Jin Kim. Give-and-take heuristic model to political redistricting problems. *Spatial Information Research*, 27(5):539–552, 2019.
- 4 David Martin. Optimizing census geography: The separation of collection and output geographies. *International Journal of Geographical Information Science*, 12(7):673–685, 1998.
- 5 Stan Openshaw. A million or so correlation coefficients: Three experiments on the modifiable areal unit problem. In *Statistical Applications in the Spatial Science*, pages 127–144. Pion, 1979.
- 6 Stan Openshaw. *The Modifiable Areal Unit Problem*. Geo Books, Norwich, 1983.
- 7 Stan Openshaw and RS Baxter. Algorithm 3: A procedure to generate pseudo-random aggregations of n zones into m zones, where m is less than n . *Environment and Planning A*, 9(12):1423–1428, 1977.
- 8 Stan Openshaw and Liang Rao. Algorithms for reengineering 1991 census geography. *Environment and Planning A*, 27(3):425–446, 1995.
- 9 United States Census Bureau. Glossary: Census tract, 2022. URL: https://www.census.gov/programs-surveys/geography/about/glossary.html#par_textimage_13.
- 10 Ningchuan Xiao, Peixuan Jiang, Myung Jin Kim, and Anuj Gadhav. A multistart heuristic approach to spatial aggregation problems. In *International Conference on GIScience Short Paper Proceedings*, volume 1, pages 349–351, 2016.

Agent-Based Modelling and Disease: Demonstrating the Role of Human Remains in Epidemic Outbreaks

Huixin Liu ✉ 🏠

The Bartlett Centre for Advanced Spatial Analysis, University College London, UK

Sarah Wise ✉

The Bartlett Centre for Advanced Spatial Analysis, University College London, UK

Abstract

Hemorrhagic fever viruses present a high risk to humans, given their associated high fatality rates, extensive care requirements, and few relevant vaccines. One of the most famous such viruses is the Ebola virus, which first came to international attention during an outbreak in 1976. Another is Marburg virus, cases of which are being reported in Equatorial Guinea at the time of writing. Researchers and governments all over the world share a goal in seeking effective ways to reduce or prevent the influence or spreading of such diseases. This study introduces a prototype agent-based model to explore the epidemic infectious progression of a simulated fever virus. More specifically, this work seeks to recreate the role of human remains in the progression of such an epidemic, and to help gauge the influence of different environmental conditions on this dynamic.

2012 ACM Subject Classification Computing methodologies → Modeling methodologies

Keywords and phrases Disease modelling, agent-based model, hemorrhagic fever virus, epidemiology, safe burial practices

Digital Object Identifier 10.4230/LIPIcs.GIScience.2023.48

Category Short Paper

Supplementary Material

Software: <https://github.com/Huixin-coder/Huixin--Giscience-2023.git>

Funding *Sarah Wise:* UKRI Grant MR/T02075X/1.

1 Introduction

Viral hemorrhagic fevers (VHFs) represent a growing threat to human health, even as recent events reflect the challenges and costs of widespread pandemics. The 2014-16 Ebola outbreak occurred primarily in West Africa killed over 11,000 people, with the World Health Organization reporting new outbreaks every single year; Marburg disease, too, has been detected with increasing frequency.[2] The main form of transmission for VHFs that spread from human to human are blood or body fluids from a human infected with Ebola.[5] It is known that in certain cases, human remains continue to be infectious; through unsafe handling of human remains or funeral ceremonies, people may infect others even after death. While well known to practitioners as a pillar of outbreak control, this dynamic has received less attention than living human-to-human contact in the simulation literature. This is unfortunate, as funeral customs in some of the areas where these diseases are endemic involve extensive contact between mourners and the body of the deceased; it is thought that this may have been a significant driver of certain outbreaks.[1][3] Thus, this paper will explore how adding corpse-to-human transmission influences an existing human-to-human model, developing a prototype agent-based model to explore the epidemic infectious progression of a theoretical hemorrhagic fever virus.



© Huixin Liu and Sarah Wise;
licensed under Creative Commons License CC-BY 4.0

12th International Conference on Geographic Information Science (GIScience 2023).

Editors: Roger Beecham, Jed A. Long, Dianna Smith, Qunshan Zhao, and Sarah Wise; Article No. 48; pp. 48:1–48:7

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

2 Background

The outbreak of Covid-19 prompted many researchers to turn their hand to the problem of epidemics, resulting in an explosion in the creation of agent-based models (ABMs) of disease (see for example [9]; [17]). The popularity of SIR (Susceptible, Infectious, and Recovered) models and its close cousins (those with states such as exposed, vaccinated, or immune) meant that researchers could track the development of disease in individual simulated persons. Agent-based models made it possible for researchers to vary the specific qualities of the individuals being exposed to disease, to control contact through social networks, and to impose non-pharmaceutical interventions on the world which had varying impacts on different groups (eg school closures versus general travel bans). Given the pressure to respond to the crisis, these simulations were naturally targeted at Covid-19 specifically - and perhaps therefore tended to deprioritise the role of the deceased in the spread of disease.

To take a more general example, [7] present *nosoi*, an open-source r package that offers a agent-based framework for simulating infectious disease events. Agents are removed from the *nosoi* model when they die - meaning that their bodies do not remain in the model to infect others. This appears to be a widespread practice across the discipline. Even when modelling Ebola specifically, [8] remove bodies upon death. [14] apply an SIR system dynamics model to Ebola, using Bayesian inference to calculate the flow among compartments representing different statuses; they add an extra compartment they call 'X' to allow them to track deaths more easily and vary the R_0 to reflect local care and funeral practices. The deterministic numerical simulation of [2] does include the role of funerals and the un/safe handling of infectious human remains. Finally, [11] builds upon the work of [6], with the former expanding upon the latter's basic compartmental model to apply the transmission process to a spatial agent-based model. The model of [11] takes into account the role of contact with the deceased during unsafe funerals; it is the only simulation we have been able to identify that considers the impact of human remains on transmission.

This work is focused on exploring the role of the human remains in the growth of an epidemic. We seek to demonstrate the significance of including or ignoring this process, investigating how the presence of human remains influences the infectious progression under different environmental conditions. Thus, this study utilises a simple agent-based model to present a series of counterfactuals. This method is computationally inexpensive enough to execute a large number of simulations and for us to pinpoint the exact role of the changing variables.

3 Methods

As this research aims to explore the impact of traditional burial practices on the spread of VHFs, we developed a basic simulation framework¹ using the Python Mesa module [10]. In the model, individual humans move randomly around the environment, potentially infecting those immediately around them with the theoretical VHF. Susceptible individuals may sicken and die, and their remains will eventually - but not immediately - be removed from the simulation.

In order to focus on the impact of time to interment - which we considered in the experiments in the following chapter, we generate an empty, theoretical environment which allows us to experiment without concern for confounding factors. Agents are randomly moving and interacting with each other on the grid. The default model parameters are as

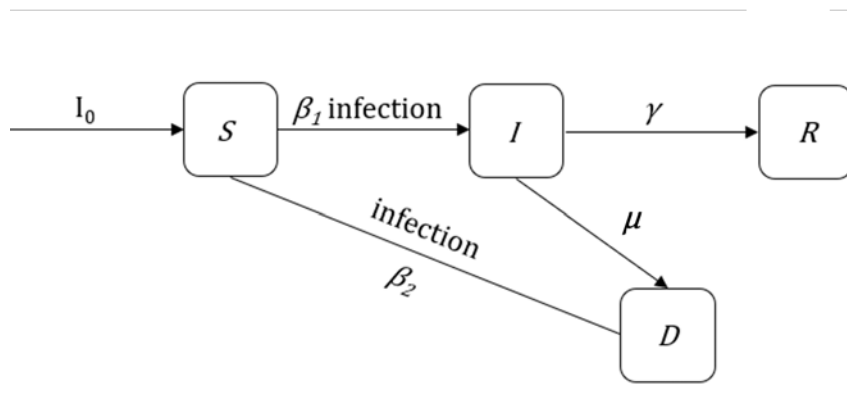
¹ Available on GitHub at <https://github.com/Huixin-coder/Huixin--Giscience-2023>

■ **Table 1** Default Model parameters.

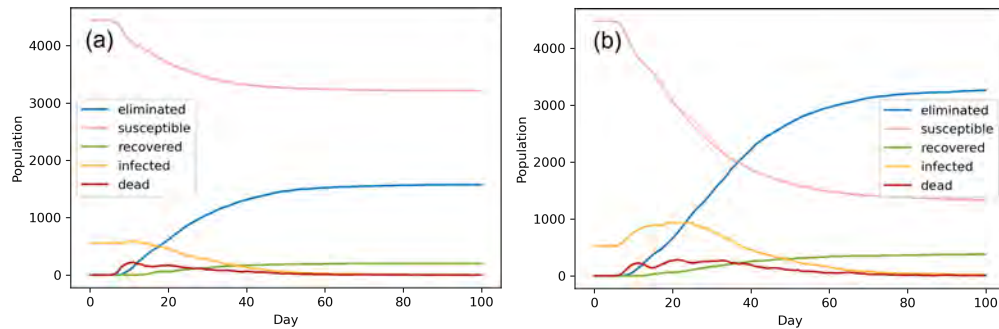
Parameter	Default Value	Reference/Assumption
Multigrid	150 x 150	—
Step	100	—
Population	5000	See the determination below
Initial infected rate	0.11	Initial infection rate is 1/9 [15]
Transmission probability	0.44	From 0.44 to 0.9 [13]
Progression period mean	8	The incubation period of 2–21 days (mean 4–10 days) [4]
Recover days mean	7	7–14 days after first symptoms [16]
Eliminated days mean	3	The virus is infectious for 7 days [12]

shown in Table 1. Similarly, the susceptible human population is held constant and will not be supplemented, as the time period being simulated is not long enough for births or natural deaths to play a significant role. The model’s step represents the a single day in the simulation.

Individuals in the population behave as visualised in Figure 1. At the beginning of the simulation, a small number of human individuals will be selected from the susceptible population S to be infected based on the initial infection rate I_0 . Susceptible individuals may acquire the infection after contact with infectious individuals (with chance β_1) or infectious human remains (with chance β_2). Infectious individuals I may recover at rate γ or die at rate μ . Deceased individuals remain temporarily in the simulation, potentially infecting others around them as controlled by the β_2 parameter. After some number of days defined by the eliminated days parameter, the human remains are removed from the environment. The parameters used in this paper are roughly based on the Ebola virus, but can easily be varied to explore other VHF.



■ **Figure 1** VHF status flowchart (SIRD): individual agents exposed to the virus may progress from susceptible (S) to infectious (I) with probability β_1 . Eventually they will experience either recovery (R) or death (D), with probabilities γ or the “death rate” respectively. Deceased agents remaining in the simulation may come into contact with other, living, susceptible agents and transmit the disease to them (with probability β_2).



■ **Figure 2** Typical sample instances of simulation results for Model 1 (a): human remains are not infectious, and Model 2 (b): human remains are infectious and can transmit the virus to living persons. Parameter values are the same in Models 1 and 2.

4 Results

In this section, we will first present a comparison of two versions of the model. In Model 1, human remains are not infectious. In Model 2, human remains **are** infectious and can transmit the virus to living persons. Building on this, we present two further experiments, exploring the impact of time to interment (called ‘eliminated days’) relative to different population densities (Experiment 1) or virus fatality rates (Experiment 2). All other parameters are held constant throughout.

4.1 Infectious versus Noninfectious Remains

Models 1 and 2 are run until the 100th timestep, at which point experimentation shows they usually equilibrate. The results do not show large deviations across either set of simulations. Figure 2 shows a comparison of the different model outcomes.

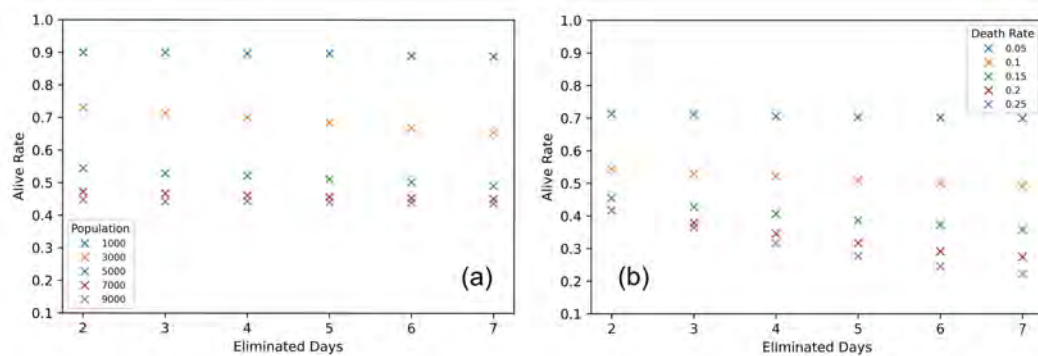
In Model 1, the R_0 typically stabilises around time 60, levelling out at 0.258, meaning that the outbreak will gradually disappear and be well controlled. Notably, such an R_0 is far from the R_0 of, say, the real-world Ebola virus which lie in the range of 1.56 to 1.9. In contrast, Model 2 with its infectious remains sees the measurements stabilise around time 70, with many more fatalities. Its R_0 value reaches about 1.6, suggesting that the disease has the potential to create an epidemic. The average final number of deceased persons are 1587.3 in Model 1, and 3161.2 in Model 2, reflecting the increased mortality associated with infectious remains.

4.2 Experiment 1: Population Density relative to Eliminated Days

As described above, Experiment 1 involves varying the population and eliminated days (2 to 7 days [12]) relative to one another, holding all other parameters as in the default model. This is meant to explore the sensitivity of the process to population density, and to better understand how significantly the timely handling of human remains impacts the spread of disease.

The model tracks the number of agents which are alive at the end of the simulation, referred to here as the “alive rate”. This is calculated by the following equation:

$$AliveRate = \frac{Susceptible + Recovered}{population}$$



■ **Figure 3** Average measures of the “Alive Rate” across 50 repetitions of (a) Experiment 1: varying population densities and number of days until human remains are eliminated from the model, and (b) Experiment 2: varying death rates and number of days until human remains are eliminated from the model.

Results are taken at the end of the 100th step. The population is set at 1000, 3000, 5000, 7000 and 9000, while the eliminated days range from 2 to 7 days. Each combination of parameters is repeated 50 times.

Figure 3(a) tracks the average “alive rate” of each combination of parameters as population and eliminated days are varied. The different population levels are clearly distinguishable, and as expected the alive rate decreases as either eliminated days or population density increases. Interestingly, the most extreme population values appear to be less affected by the speed with which remains are handled. In contrast, the sensitivity of the population of 3000 to the number of eliminated days is related to the size of the population relative to the size of the grid. The uneven distribution of alive rates relative to population size at any given value of eliminated days suggests that there may be critical points of inflection in model behaviour.

What the graph suggests is that in situations of medium population density when a susceptible person might not otherwise encounter an infectious living person, the long-term presence of infectious remains represents a noticeable peril.

4.3 Experiment 2: Change daily death rate and eliminated days

Experiment 2 holds population constant (size 5000) and instead varies the fatality of the infection relative to the eliminated days. The daily death rate increases from 0.05 to 0.25 in increments of 0.05, while the eliminated days again range from 2 to 7. Once more, each parameter combination is run 50 times.

Figure 3 (b) shows the relationship between the daily death rate and eliminated days as defined by the average alive rate. Again, as expected the alive rate decreases with the increase in number of eliminated days, regardless of daily death rate level. In certain situations, infections are known to “burn themselves out” by killing off hosts before a virus has the opportunity to spread to new hosts. If human remains are infectious, however, the highly virulent strains of disease are still able to spread, especially when these deceased hosts remain in the environment.

5 Discussion and Conclusion

This article demonstrates a simple example of how improperly handled infectious human remains can propagate and worsen epidemics. Many extant modelling frameworks remove deceased agents immediately; our goal is to show the impact that such a modelling choice may have. There are of course often reasons for such coding decisions. For example, in extremely large-scale models being run on suboptimal hardware setups, recovering memory may be a priority. However, we would caution against adopting such a framework without careful consideration. At a minimum, modellers should be aware of the impact such decisions have on the ultimate course of an epidemic.

Simulation as a tool showed a great deal of promise during the recent Covid-19 pandemic. It is crucial, however, that researchers ensure that models not sacrifice essential functionality in the name of parsimony. This paper presents a simple example drawn from a well-known principle of infectious suppression. It is important that modellers engage proactively with subject matter experts to ensure that we incorporate such dynamics into our work in the future.

References



- 1 Sharon Alane Abramowitz. Epidemics (Especially Ebola). *Annual Review of Anthropology*, 2017. doi:10.1146/annurev-anthro-102116-041616.
- 2 Tsanou Berge, Jean M.-S. Lubuma, G.M. Moremedi, G. M. Moremedi, Neil Kenneth Morris, Neil Kenneth Morris, and R. Kondera-Shava. A simple mathematical model for Ebola in Africa. *Journal of Biological Dynamics*, 2017. doi:10.1080/17513758.2016.1229817.
- 3 James Fairhead. The significance of death, funerals and the after-life in Ebola-hit Sierra Leone, Guinea and Liberia: Anthropological insights into infection and social resistance. Technical report, Institute for Development Studies, University of Nairobi, 2014. URL: <https://opendocs.ids.ac.uk/opendocs/handle/20.500.12413/4727>.
- 4 Heinz Feldmann and Thomas W. Geisbert. Ebola haemorrhagic fever. *The Lancet*, 2011. doi:10.1016/s0140-6736(10)60667-8.
- 5 Centers for Disease Control and Prevention (CDC). What are VHF's?, 2021. URL: <https://www.cdc.gov/vhf/about.html>.
- 6 Judith Legrand, Rebecca F. Grais, Pierre-Yves Boëlle, Alain-Jacques Valleron, Antoine Flahault, and Antoine Flahault. Understanding the dynamics of ebola epidemics. *Epidemiology and Infection*, 2007. doi:10.1017/s0950268806007217.
- 7 Sebastian Lequime, Paul Bastide, Simon Dellicour, Philippe Lemey, and Guy Baele. Nosoi: A stochastic agent-based transmission chain simulation framework in R. *Methods in Ecology and Evolution*, 2020. doi:10.1111/2041-210x.13422.
- 8 Xueping Li and Shima Mohebbi. Modeling Diffusion of Epidemic Diseases via Agent-based Simulation. *IIE Annual Conference Proceedings*, pages 2156–2162, 2015. Copyright - Copyright Institute of Industrial Engineers-Publisher 2015; Document feature - Diagrams; Tables; Graphs; ; Last updated - 2022-11-13. URL: <https://www.proquest.com/scholarly-journals/modeling-diffusion-epidemic-diseases-via-agent/docview/1792022743/se-2>.
- 9 Fabian Lorig, Emil Johansson, and Paul Davidsson. Agent-Based Social Simulation of the Covid-19 Pandemic: A Systematic Review. *Journal of Artificial Societies and Social Simulation*, 24(3):5, 2021. doi:10.18564/jasss.4601.
- 10 David Masad and Jacqueline L. Kazil. Mesa: An agent-based modeling framework. *SciPy*, 2015. doi:10.25080/majora-7b98e3ed-009.
- 11 Stefano Merler, Marco Ajelli, Laura Fumanelli, Marcelo F. C. Gomes, Ana Pastore y Piontti, Luca Rossi, Dennis L. Chao, Ira M. Longini, M. Elizabeth Halloran, and Alessandro Vespignani.

- Spatio-temporal spread of the Ebola 2014 outbreak in Liberia and the effectiveness of non-pharmaceutical interventions: a computational modelling analysis. *Lancet Infectious Diseases*, 2015. doi:10.1016/s1473-3099(14)71074-6.
- 12 Joseph Prescott, Trenton Bushmaker, Robert S. Fischer, Robert J. Fischer, Robert J. Fischer, Kerri L. Miazgowicz, Seth D. Judson, and Vincent J. Munster. Postmortem stability of Ebola virus. *Emerging Infectious Diseases*, 2015. doi:10.3201/eid2105.150041.
 - 13 Suresh Rewar and Dashrath Mirdha. Transmission of Ebola Virus Disease: An Overview. *Annals of Global Health*, 2015. doi:10.1016/j.aogh.2015.02.005.
 - 14 Jeffrey Shaman, Wan Yang, and Sasikiran Kandula. Inference and Forecast of the Current West African Ebola Outbreak in Guinea, Sierra Leone and Liberia. *PLOS Currents*, 2014. doi:10.1371/currents.outbreaks.3408774290b1a0f2dd7cae877c8b8ff6.
 - 15 Constantinos I. Siettos, Cleo G. Anastassopoulou, Lucia Russo, Christos Grigoras, and Eleftherios Mylonakis. Modeling the 2014 Ebola Virus Epidemic – Agent-Based Simulations, Temporal Analysis and Future Predictions for Liberia and Sierra Leone. *PLOS Currents*, 2015. doi:10.1371/currents.outbreaks.8d5984114855fc425e699e1a18cdc6c9.
 - 16 G. Thomas Strickland. Hunter's Tropical Medicine and Emerging Infectious Diseases. *Revista do Instituto de Medicina Tropical de São Paulo*, 2019. doi:10.1590/s0036-46652001000200018.
 - 17 Jing Tang, Sukrit Vinayavekhin, Manapat Weeramongkolkul, Chanakan Suksanon, Kantapat Pattarapremcharoen, Sasinat Thiwathittayanuphap, and Natt Leelawat. Agent-Based Simulation and Modeling of COVID-19 Pandemic: A Bibliometric Analysis. *Journal of Disaster Research*, 17(1):93–102, January 2022. doi:10.20965/jdr.2022.p0093.

How Does Travel Environment Affect Mood? A Study Using Geographic Ecological Momentary Assessment in the UK

Milad Malekzadeh¹  

Western University, London, Canada

Darja Reuschke  

University of Southampton, UK

Jed A. Long  

Western University, London, Canada

Abstract

Daily travel is a large part of life, and it is widely believed that our mood can be affected by the environment in which travel takes place. In this study, we investigate how environmental factors affect mood while performing daily travel activities using an app-based geographic ecological momentary assessment study. Our study (the WorkAndHome study) involved over 1000 participants tracked using a bespoke GPS mobile phone app in three cities (Birmingham, Leeds, and Brighton and Hove, UK) At the end of trips (i.e., when a stop in the GPS data was detected) we pushed a survey to participants asking them to score their current happiness and stress levels on a 7-point Likert scale. We combined individual GPS data with environmental data on green and blue spaces and weather conditions. We found that green and blue space availability and weather variables, such as daytime, apparent temperature, and visibility, significantly affect our happiness levels at the end of trips. While these weather factors were also significant predictors of stress level, availability of green and blue space was not. The results of this study provide fine-scale evidence from direct surveys about the associations between environment and weather and our moods when performing daily travel activities.

2012 ACM Subject Classification Information systems → Geographic information systems

Keywords and phrases GEMA, GPS Tracking, Green and Blue Spaces

Digital Object Identifier 10.4230/LIPIcs.GIScience.2023.49

Category Short Paper

1 Introduction

It is well documented that a variety of trip attributes; such as mode of transportation [5], the duration of the trips [17], the type of activity [7], and whether we are travelling alone or not, can affect our mood during and following trips. Less is known about how the environment where trips occur influences mood. Previous evidence supports that trips occurring in greenspaces are associated with greater happiness levels [23]. Further, it is believed that the environmental features where we conduct our trips can significantly influence our mood [5]. To study the effect of environmental factors on our mood, we need to capture individuals' immediate experiences during and/or immediately following trips. Geographic ecological momentary assessment (GEMA) therefore represents an ideal method to track real-time data on how individuals feel. Previous studies have successfully employed GEMA methods to investigate human exposure and response of the environment on people using GPS-enabled

¹ Corresponding author



© Milad Malekzadeh, Darja Reuschke, and Jed A. Long;
licensed under Creative Commons License CC-BY 4.0

12th International Conference on Geographic Information Science (GIScience 2023).

Editors: Roger Beecham, Jed A. Long, Dianna Smith, Qunshan Zhao, and Sarah Wise; Article No. 49; pp. 49:1–49:6

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

apps on mobile phones [18]. In this study, we use GEMA to collect targeted information on individual mood (happiness and stress) and investigate how mood is associated with environmental factors such as green and blue spaces and weather conditions.

2 Data and Methods

We used the WORKANDHOME dataset [11, 21] comprising mobile-phone based GPS data for 1029 participants in three UK cities (Brighton and Hove, Leeds, and Birmingham). This data were collected in two sampling periods: Oct 2018 to May 2019 (Leeds, Brighton & Hove) and Sep 2019 to Apr 2020 (Birmingham). We tracked each participant’s movement (with their consent) and pushed a GEMA survey corresponding to any trip endpoint detected by the app. In the GEMA survey, we asked a set of 6 questions (Table 1), including questions about mood (happiness, stress, and enjoyment) on a 7-point Likert scale. Along with the GPS data and GEMA survey, we collected detailed socio-demographic information on each participant through a telephone-based survey administered prior to installing our mobile phone app. In this study, we incorporated the self-reported variables on gender, age, and having a health issue limiting mobility.

■ **Table 1** GEMA survey questions and their possible responses.

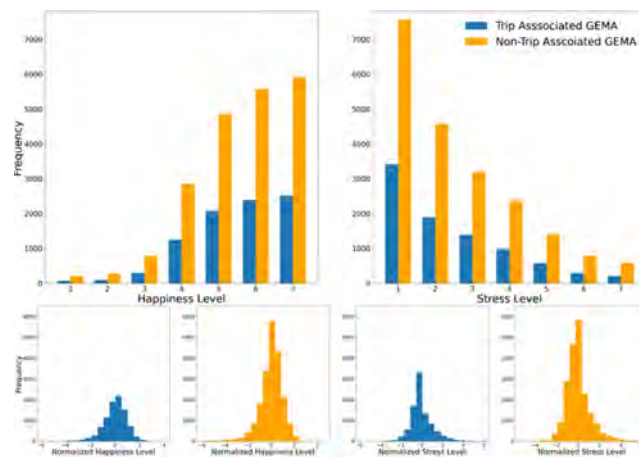
Questions	Responses	Questions	Responses
Where are you?	Work, home, other	How happy are you?	1-7 (the least to the most)
Whom are you with?	Alone, not alone	How stressed are you?	1-7 (the least to the most)
What activity are you involved in?	Work, housework, leisure, eating, other	How much are you enjoying?	1-7 (the least to the most)

Using methods detailed in [11] we derived trips from individual’s raw GPS tracking data. In total, we extracted 31743 trips. However, not all trips have a completed GEMA response at the end, and we kept only those trips where the GEMA survey was completed within 1 hour of the trip end time. After filtering out trips with successful surveys, we had a dataset of 8654 trips from 657 different participants. We used Meteorological Office Integrated Data Archive System (MIDAS) data to assign weather attributes to each trip in our study. MIDAS is a comprehensive weather database managed by the UK’s national weather service [14, 15]. Here we used hourly data for rainfall and other weather attributes: air temperature, air pressure, wet bulb temperature, wind speed, and horizontal visibility. Previous research has demonstrated that apparent temperature is a useful variable for capturing how human beings experience weather and therefore we calculated the apparent temperature (in Centigrade) [22]. We used the UK Centre for Ecology and Hydrology (UKCEH) land cover dataset to extract information on green and blue spaces [13]. UKCEH uses Sentinel-2 Seasonal Composite Images reflecting the median reflectance for each season. The land cover dataset is comprised of 21 classes of land cover. We merged 11 green-related classes as green space and two blue-related classes as blue space. We calculated the area of green and blue space present in a buffer of 50 meters around each trip’s GPS data. The area of green and blue space within each trip was divided by the area of the 50 m buffer to give a numerical proxy (between 0 and 1) for how much of a trip was in areas where green and/or blue space was present. As previous literature has reported [12], transport mode can significantly affect our mood. Following [24] we employed a Fuzzy Logic system to detect the mode of transportation. We used 6 transportation mode categories: walk, run, bike, bus, train, and car. We used four parameters: median speed, standard deviation speed, proximity to bus routes, and proximity to train routes. Incorporating four parameters enabled us to distinguish between modes that are similar in one aspect but different in the other. For example, bus and car might

have the same median speed, but their proximity to bus routes is different; consequently, our fuzzy system differentiates these two from each other. We employed min-max operation (minimum value in each parameter and maximum value between all mode categories) to identify each mode of transportation. To employ public transit in our model, we used the Open Street Map (OSM) dataset to extract train and bus routes of any kind. We used a linear mixed-effect regression model to account for participants having multiple trips as a random effect. We considered two GEMA response variables (happiness and stress) measured on a 7-point Likert scale. Prior to analysing the data, we adjusted each participants GEMA scores for happiness and stress by subtracting the mean response for each individual across all GEMA surveys (including those GEMA surveys not associated with a trip) from each response.

3 Results

More than two third (69%) of GEMA surveys were not associated with a trip. This provides a comprehensive assessment of happiness and stress levels in various contexts. We observe no significant difference in happiness and stress levels between trip and non-trip GEMA surveys responses (Figure 1). The average and standard deviation of happiness levels for trip GEMA surveys are 4.55 and 1.26, and for non-trip GEMA surveys are 4.55 and 1.30. Similarly, the average and standard deviation of stress levels for trip GEMA surveys are 1.47 and 1.58, and for non-trip GEMA surveys are 1.52 and 1.63.



■ **Figure 1** Distributions of raw self-reported happiness and stress levels and their normalized values.

Higher green/blue spaces were found to be positively associated with Happiness scores (Table 2) but showed no significant association with stress level. We found daytime had a negative association with happiness level; meaning individuals had higher happiness scores at night than during the day. Similarly, we found that daytime was positively associated with stress level. Apparent temperature was positively associated with happiness and negatively associated with stress. Rainfall showed no significant association with either of happiness or stress. Travel mode was not found to have an overall significant impact on GEMA happiness or stress scores, with the exception of bus travel, which was negatively associated with happiness (Table 2). Destination type also did not significantly influence observed happiness or stress scores. Trip duration was positively associated with stress level (but not happiness); whereas trip length was negatively associated with stress (but not happiness). Housework, leisure, and other activities were not significantly associated with happiness or stress levels

49:4 How Does Travel Environment Affect Mood?

compared to work as the reference category. Travelling with someone (vs. alone) was not found to be associated with happiness or stress. We found no associations between individual factors (age, gender, whether or not individuals self-report a health issue that limits mobility) and happiness or stress.

■ **Table 2** Results of linear mixed-effect regression models of happiness and stress level.

Predictors	Happiness		Stress	
	Estimates	p	Estimates	p
(Intercept)	-0.27	0.177	0.17	0.445
Green-Blue Spaces	0.40	0.032	-0.20	0.317
Daytime	-0.26	<0.001	0.23	<0.001
Apparent Temperature	0.24	0.006	-0.21	0.030
Visibility	0.43	<0.001	-0.22	0.021
Rain	0.02	0.508	-0.05	0.154
Travel Mode - Walk	-0.01	0.840	-0.08	0.124
Travel Mode - Run	0.07	0.593	-0.12	0.378
Travel Mode - Bike	-0.00	0.951	-0.06	0.465
Travel Mode - Bus	-0.17	0.020	0.10	0.200
Travel Mode - Train	-0.20	0.376	0.21	0.411
Travel Mode - Car	-0.08	0.067	0.02	0.718
Destination Type [RC: Home]				
Work	0.02	0.679	-0.04	0.281
Other	-0.02	0.568	0.01	0.798
Duration	-0.43	0.072	0.57	0.030
Length	0.23	0.336	-0.55	0.034
Activity Type [RC: Work]				
Housework	0.05	0.329	0.00	0.957
Leisure	0.02	0.598	0.05	0.144
Eating	-0.36	0.058	0.16	0.445
Other	0.12	0.447	-0.18	0.283
Presence of People - Not Alone [RC: Alone]	-0.03	0.254	-0.00	0.961
Health and Mobility Issue - Yes [RC: No]	-0.03	0.676	0.08	0.259
Gender - Male [RC: Female]	-0.00	0.995	0.00	0.989
Age [RC: 18-24]				
25-34	0.04	0.433	-0.08	0.152
35-44	0.02	0.703	-0.08	0.150
45-54	-0.02	0.716	-0.04	0.525
55-64	-0.01	0.804	-0.02	0.762
σ^2	0.98		1.18	
τ_{00}	0.01		0.02	
Marginal R2 / Conditional R2	0.022/0.034		0.014/0.028	

RC: Reference Category. Bold number: significant association.

4 Discussion and Conclusion

In line with the existing literature, we incorporated environmental and weather factors into our study, as they have been commonly studied in relation to self-reported happiness and stress levels during and after trips. While previous studies have demonstrated that spending

time in green and blue spaces may reduce stress levels [4], we found that daily travel through these spaces is not significantly associated with individuals' stress levels. An important difference in our study is that we are measuring the proportion of the trip through green-blue spaces by area rather than measuring time spent in those spaces, which may differentiate what we have found with previous studies. It is interesting that our results support a positive association between the amount of green-blue spaces experienced during trips and happiness levels, which is a similar effect as to when individuals spend time in these spaces [10].

We found daytime was negatively associated with happiness level and positively associated with stress level. One reason for this might be that it is estimated that 96% of workers are daytime workers [1], and as work is recognized as a significant source of stress [2], it is not unexpected to find daytime a positive predictor of stress level. Similarly we found a positive association between apparent temperature and happiness, but a negative association with stress. Previous studies have found that individuals spend more time on leisure and fun activities [9], and have a better mood [8] during warmer days and seasons which might explain this relationship. We also found horizontal visibility to have the same relationships with happiness and stress, respectively. It has been previously identified that foggy weather and a high level of humidity can negatively affect individuals' moods [25]. Moreover, another reason for this might be that individuals feel safer travelling when visibility is greater.

It is also interesting that we found all transport modes to be non-significant predictors of happiness and stress, with the exception of travel by bus which was negatively associated with happiness. Previous research has reported active transportation and private transportation may positively affect our mood [6]. We limited evidence on the impact of these individual factors, while previous studies have identified significant associations between individual factors and mood during trips [20].

It is worth noting that mood is a complex response which is difficult to capture in survey data, and therefore often difficult to measure [19]. In our study, we limited our analysis to investigating the role of daily travel and the surrounding environment (i.e., weather and green/blue spaces) on individuals' moods. Numerous factors, including individual genetics and personal characteristics [3], and interpersonal connections [16] can affect individuals' moods. We tried to control for this by adjusting the happiness and stress levels by individuals' average scores. However, there are many other varying factors that we cannot control for. Therefore, it is likely that the complexity of individual happiness and stress levels may limit the explanatory power of our models (as observed here, overall model fit was low ($R^2 < 5\%$)).


In conclusion, we found that travel environment (such as the presence of green and blue spaces and weather characteristics) was significantly associated with mood (happiness and stress). These results highlight the importance of green and blue spaces in our travel environment. Increasing green and blue spaces along travel routes, especially in urban spaces, can potentially improve citizens' travel-related well-being.

References

- 1 Office for national statistics, 2023. URL: <https://www.ons.gov.uk>.
- 2 American Psychological Association. Stress and decision-making during the pandemic, 2021.
- 3 L. Bevilacqua and D. Goldman. Genetics of emotion. *Trends in Cognitive Sciences*, 15(9):401–408, 2011.
- 4 Sjerp De Vries, Margreet Ten Have, Saskia van Dorsselaer, Manja van Wezep, Tia Hermans, and Ron de Graaf. Local availability of green and blue space and prevalence of common mental disorders in the netherlands. *BJPsych open*, 2(6):366–372, 2016.

- 5 A. Duarte, C. Garcia, G. Giannarakis, S. Limão, A. Polydoropoulou, and N. Litinas. New approaches in transportation planning: happiness and transport economics. *NETNOMICS: Economic Research and Electronic Networking*, 11:5–32, 2010.
- 6 L. Eriksson, M. Friman, and T. Gärling. Perceived attributes of bus and car mediating satisfaction with the work commute. *Transportation Research Part A: Policy and Practice*, 47:87–96, 2013.
- 7 D. Ettema, M. Friman, T. Gärling, L.E. Olsson, and S. Fujii. How in-vehicle activities affect work commuters' satisfaction with public transport. *Journal of Transport Geography*, 24:215–222, 2012.
- 8 M.C. Keller, B.L. Fredrickson, O. Ybarra, S. Côté, K. Johnson, J. Mikels, A. Conway, and T. Wager. A warm heart and a clear head: The contingent effects of weather on mood and cognition. *Psychological Science*, 16(9):724–731, 2005.
- 9 Y. Kim and R. Brown. Effect of meteorological conditions on leisure walking: a time series analysis and the application of outdoor thermal comfort indexes. *International Journal of Biometeorology*, 66(6):1109–1123, 2022.
- 10 M.C. Kondo, M. Triguero-Mas, D. Donaire-Gonzalez, E. Seto, A. Valentín, G. Hurst, G. Carrasco-Turigas, D. Masterson, A. Ambròs, and N. Ellis. Momentary mood response to natural outdoor environments in four european cities. *Environment International*, 134:105237, 2020.
- 11 J. Long and D. Reuschke. Daily mobility patterns of small business owners and homeworkers in post-industrial cities. *Computers, Environment and Urban Systems*, 85, 2021.
- 12 P.L. Mokhtarian and R.M. Pendyala. Travel satisfaction and well-being. *Quality of Life and Daily Travel*, pages 17–39, 2018.
- 13 R.D. Morton, C.G. Marston, A.W. O'Neil, and C.S. Rowland. Land cover map 2019 (20m classified pixels, gb. *NERC Environmental Information Data Centre*, 2020.
- 14 Met Office, 2006.
- 15 Met Office. Midas: Uk hourly weather observation data, 2006.
- 16 Y. Ogihara and Y. Uchida. Does individualism bring happiness? negative effects of individualism on interpersonal relationships and happiness. *Frontiers in Psychology*, 5:135, 2014.
- 17 L.E. Olsson, T. Gärling, D. Ettema, M. Friman, and S. Fujii. Happiness and satisfaction with work commute. *Social Indicators Research*, 111:255–263, 2013.
- 18 E.M. Parrish, C.A. Depp, R.C. Moore, P.D. Harvey, T. Mikhael, J. Holden, J. Swendsen, and E. Granholm. Emotional determinants of life-space through gps and ecological momentary assessment in schizophrenia: what gets people out of the house? *Schizophrenia Research*, 224:67–73, 2020.
- 19 R. Plutchik. The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American Scientist*, 89(4):344–350, 2001.
- 20 S. Raveau, A. Ghorpade, F. Zhao, M. Abou-Zeid, C. Zegras, and M. Ben-Akiva. Smartphone-based survey for real-time and retrospective happiness related to travel and activities. *Transportation Research Record*, 2566(1):102–110, 2016.
- 21 D. Reuschke. Workandhome, 2015-10. URL: <http://workandhome.ac.uk/>.
- 22 R.G. Steadman. Norms of apparent temperature in australia. *Aust. Met. Mag*, 43:1–16, 1994.
- 23 R. Wang, Z. Feng, J. Pearce, S. Zhou, L. Zhang, and Y. Liu. Dynamic greenspace exposure and residents' mental health in guangzhou, china: From over-head to eye-level perspective, from quantity to quality. *Landscape and Urban Planning*, 215:104230, 2021.
- 24 C. Xu, M. Ji, W. Chen, and Z. Zhang. Identifying travel mode from gps trajectories through fuzzy pattern recognition. In *Seventh International Conference on Fuzzy Systems and Knowledge Discovery*, volume 2, pages 889–893, 2010.
- 25 I. Čelić, S. Živanović, and N. Pavlović. The effects of weather conditions on the health of people living in urban and rural environments. *Economics of Agriculture*, 66(1):63–76, 2019.

Calibration in a Data Sparse Environment: How Many Cases Did We Miss?

Robert Manning Smith ✉ 

The Bartlett Centre for Advanced Spatial Analysis, University College London, UK

Sarah Wise ✉ 

The Bartlett Centre for Advanced Spatial Analysis, University College London, UK

Sophie Ayling ✉ 

The Bartlett Centre for Advanced Spatial Analysis, University College London, UK

Abstract

Reported case numbers in the COVID-19 pandemic are assumed in many countries to have underestimated the true prevalence of the disease. Deficits in reporting may have been particularly great in countries with limited testing capability and restrictive testing policies. Simultaneously, some models have been accused of over-reporting the scale of the pandemic. At a time when modeling consortia around the world are turning to the lessons learnt from pandemic modelling, we present an example of simulating testing as well as the spread of disease. In particular, we factor in the amount and nature of testing that was carried out in the first wave of the COVID-19 pandemic (March - September 2020), calibrating our spatial Agent Based Model (ABM) model to the reported case numbers in Zimbabwe.

2012 ACM Subject Classification Computing methodologies → Modeling methodologies

Keywords and phrases Agent Based Modelling, Infectious Disease Modelling, COVID-19, Zimbabwe, SARS-CoV-2, calibration

Digital Object Identifier 10.4230/LIPIcs.GIScience.2023.50

Category Short Paper

Funding *Robert Manning Smith*: UKRI Grant MR/T02075X/1.

Sarah Wise: UKRI Grant MR/T02075X/1.

Sophie Ayling: UBEL-Doctoral Training Partnership ES/P000592/1.

1 Introduction

From the early stages of the COVID-19 pandemic, there have been initiatives to estimate the true scale and impact of the epidemic in terms of cases, hospitalizations and deaths across different countries around the world. Starting with the World Health Organization [23], a number of other data trackers sprung up (e.g. [15, 21, 12] or the more policy-focused [3]). These trackers fed into disease models which sought to predict the future spread of disease. Agent-based models (ABMs) became popular, especially as researchers sought more granular dimensions to population characteristics and scenario modelling (see [4, 16, 10]).

During the pandemic, criticisms were levelled at modellers in the public eye that the model forecasts did not reflect the number of cases that were reported in the media [2]. Certain studies suggested that the cases detected and reported were substantially under-reporting the true magnitude of the epidemic. In different contexts, researchers estimated that true case numbers might outstrip reported case numbers by a factor of between 5 and 20 ([19]. What accounts for this discrepancy?

In this paper, we attempt to recreate these “hidden” cases, taking as a case study Zimbabwe. We endeavour to replicate the true reported case numbers by layering a simulated testing process on top of our existing model of disease. The work presented in this paper



© Robert Manning Smith, Sarah Wise, and Sophie Ayling;
licensed under Creative Commons License CC-BY 4.0

12th International Conference on Geographic Information Science (GIScience 2023).

Editors: Roger Beecham, Jed A. Long, Dianna Smith, Qunshan Zhao, and Sarah Wise; Article No. 50; pp. 50:1–50:7

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

incorporates the available data on Zimbabwe's pandemic response policy, testing, and reported cases. The following sections will address some relevant background for this question (Section 2) before presenting the modelling framework and data used to inform it (Section 3). The results of the applied model will be presented (Section 4) and contextualised (Section 5).

2 Background

This section will present motivating context for understanding reported cases of disease as well as Zimbabwe's handling of the COVID-19.

2.1 Understanding reported cases

ABMs experienced an explosion in popularity as a result of the COVID-19 pandemic. The ways in which researchers sought to understand how their simulations related to reported cases varied. Some modelers have made efforts to either a-priori include an understanding of testing, resulting in only a proportion of cases being detected, or to somehow back-calibrate to reported data. For example, the US based Institute of Disease Modelling's model Covasim [16] added a parameter to incorporate testing. Others tried to compare actual and simulated hospital admissions [17] or to calibrate their models on diagnosis versus mortality rates [14].

In many Low and Middle Income Country (LMICs) contexts, where testing capacities were often more limited, these underestimates on reported case numbers are likely to have been at least as high as those in High Income Countries (HICs). Many have proactively attempted to mitigate this: for example, in Kenya, researchers used a combination of serological and PCR test data to calibrate their work for this reason [20]. Research seems to support the idea that true cases were undercounted: in Kazakhstan, researchers used death and the Case Fatality Ratios (CFR) to attempt to backcast true case numbers from July 2020 to May 2021 of the pandemic in that country [22]. The authors of the study asserted that official cases reported undercounted the number of infections by at least 60%. A similar situation was reported in various African countries [6], where serological surveys also retrospectively appeared to reveal a much higher prevalence of those who had developed SARS-CoV-2 antibodies in the population than the reported case statistics would appear to show. For example across 3 high density suburbs in Harare, Zimbabwe researchers found that the seroprevalence was at 19% in 2020 and 53% in 2021, with almost half of the participants who tested positive reporting no symptoms in the preceding six months [11]. With this background, it is useful to explore further the specific case of Zimbabwe.

2.2 The case of Zimbabwe

Zimbabwean authorities acted very quickly after the first case was detected in their country on 20th March 2020 [5]. They launched the country's Preparedness and Response Plan for Coronavirus the very next day. However, during this initial period testing was very limited. Large scale rapid diagnostic testing did not become available till September 11th, 2020 [13]. As of 27th June 2020, Zimbabwe had 567 confirmed SARS-CoV-2 cases [21]. Eighty-two percent of these were returning residents and 18% were the result of local transmission. The testing was heavily skewed towards returnees despite a comprehensive testing strategy [18]. For those tests that were conducted, there were also logistical issues in transporting samples to the few available testing centers (see [7]) further confounding the picture. Thus, despite proactive measures by leadership, it is likely that cases in Zimbabwe were substantially underreported.

With this understanding of the need for simulation which can calibrate against systematically underreported data, we proceed to a description of the method we adopt in the rest of this paper.

3 Methodology

This model is an extension of work documented in [24], based on simulation available as an open-source project available online¹. To briefly review the simulation framework, we constructed a spatial agent-based model (ABM) simulating the spread of SARS-CoV-2 in Zimbabwe with district level dis-aggregation in movement patterns for individual agents in the model. Default model values are taken from [16], which in turn draws upon [10].

In this paper, we introduce the incorporation of a testing regime into the model to enable us to measure both cases that *exist* and cases that have been *detected* in the population.

3.1 The testing regime

The modelled testing regime sits on top of our existing simulation of the spread of the virus. In the testing regime, a number of tests are distributed amongst the population each day. Individuals who exhibit symptoms of SARS-CoV-2 are eligible for testing. The symptoms of SARS-CoV-2 - such as a continuous cough or fever - are common to many other infections; thus we take into account that people without SARS-CoV-2 will present for testing. To simulate the allocation of tests to those without the infection, we generate a number of people with “spurious” SARS-CoV-2 symptoms. These symptoms will last for 7 days before subsiding. A person will seek a test only once. This process is based upon the work of numerous contextual studies (see [7, 6, 13, 8, 9, 5]).

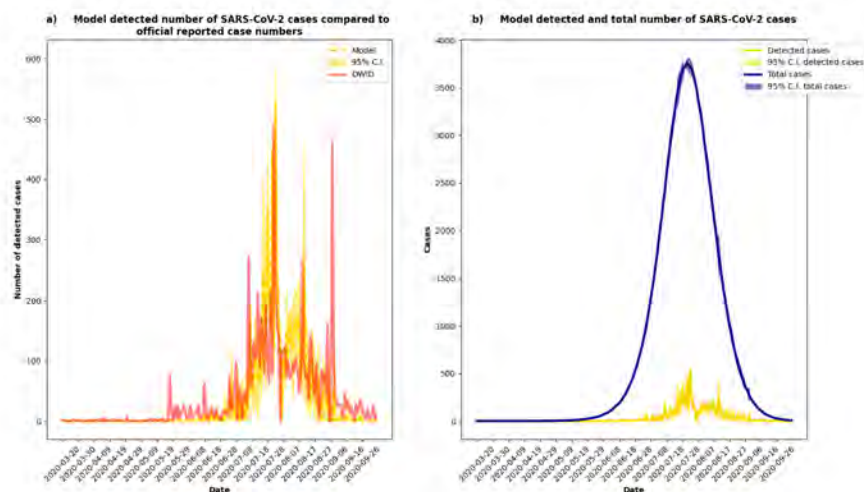
Two factors will necessarily influence the number of detected cases beyond the actual underlying number of cases: the number of tests administered per day and the number of people with SARS-CoV-2 infections who are tested. The number of tests given out each day is a set number taken from the government’s reported numbers [21]. Because the number of tests distributed daily was not available to us, we calculated the number of tests performed each day from the reported number of cases and the percent of tests that were positive as per [21]. The total number of tests administered each day were then scaled to match the models population size. The number of people with SARS-CoV-2 who are tested remains an unknown; false positives and negatives make it impossible to objectively determine this. Thus, we explore different possibilities in the results section.

3.2 Movement

One key feature of the model is the movement of individuals between districts. As we wanted to compare our test results to real reported case data, it was important to ensure that lockdowns and their consequent lower mobility levels were incorporated into the simulation.

The model calculates the likelihood of any agent moving between districts based on a number of different factors: their economic status, the day of the week, and the baseline likelihood of moving between their current district and another. That last factor is represented in the model by an origin-destination (OD) matrix, which draws from Call Detail Records (CDR) provided by the largest mobile phone service provider in the country. The raw data

¹ see <https://github.com/dime-worldbank/Disease-Modelling-SSA>



■ **Figure 1** a) The model’s number of SARS-CoV-2 cases detected through the testing regime compared to the official reported case number, taken from Our World in Data. b) The model’s detected number of SARS-CoV-2 and the total number of the model’s predicted cases.

(to which this study did not have access) covered the period February 1–June 30, 2020. At the dis-aggregated level, it contains data on 1900 towers to include 8.1 billion observations across each of the country’s 60 districts. The World Bank research team which handled this data partitioned it into two periods: the first from February 2 to March 14 (prior to the first Level 4 lockdown), and the second from March 15 – June 2020. By extracting the inter-district movements for these two time periods into separate OD matrices, they created patterns of travel representative of both normal and lockdown conditions.

Thus, in order to ensure that our simulated individuals were moving correctly, we applied a “lockdown” in the simulation by drawing the movement of individuals from a distribution defined by either the pre- or lockdown OD matrices. The simulation imposes a level 4 lockdown on the 30th of March, with reduced movement; we then revert back to the pre-lockdown levels of interdistrict travel on the 17th of May, when the imposed restrictions on intercity travel were removed as part of Level 2 measures (as per [5]).

4 Results and Discussion

Each instantiation of the model was run for 200 simulated days; our model start date and testing routine coincides with the start of the case reporting from Zimbabwe from the 20th of March 2020. The simulated population is based on a 5% sample of the 2012 Zimbabwe Census was taken from IPUMS International [1], allowing us to incorporate realistic distributions of age, sex, economic status, and household composition.

We performed a parameter grid search to calibrate the models’ number of detected cases to those reported. We paired combinations of the infection transmission parameter, β , to the rate in which a person will develop spurious symptoms, γ . The total error in the number of detected cases in each parameter combination was assessed and models were selected to minimize the total error. Initially, SARS-CoV-2 testing in Zimbabwe was limited to points of entry (functionally, districts with an official boarder crossing, airport or train station).

Within our parameter grid search, the parameter combination which resulted in model runs that most closely fit the true reported case data came when $\beta = 0.128$ and $\gamma = 0.0875$. The simulated reported cases are shown in Figure 1a, with a 95% confidence interval indicating the variation among runs. Figure 1b demonstrates the total number of simulated cases in the same model, demonstrating the significant number of cases missed as a result of a limited testing regime. During the simulation period, the model's daily detected number of cases peaked at 531, whereas the peak number of both undetected and detected cases was 3763.

Our methodology of filtering the model's simulated cases through a simulated testing regime allowed us to closely match the reported case numbers. Over the course of the simulation, the model generated a total of 153,807 cases, yet the simulated testing documented only 6892 cases. Thus, only 5% of the model's "true" cases were discovered by the testing regime. Other modelling studies have found similar discrepancies in the detected and total cumulative number of cases estimated (see for example [19]).

5 Conclusion

The results of this paper are dependent on the outcome of the model's calibration and a number of assumptions made. For example, one relevant assumption is the number of cases distributed in the population at the beginning of the simulation. Initially, we created a single infection in a 25% scale size population (equivalent to four initial cases, once scaling is taken into account). A single initial infection was chosen to represent the single initial case reported on the 20th of March. It may be that more cases existed in Zimbabwe at the time; however, in hindsight it would be impossible to establish the exact number. Seeding more infections initially would result in an increased number of cases overall. Future work might explore the sensitivity of the epidemic to the number of initial cases as well as the parameters β and γ .

Broadly, this work contributes to the discussion around disease forecasting and prediction. As described above, many people were skeptical of the apparent "overprediction" of cases of SARS-CoV-2 cases. Our results show a clear example of how the results of such simulations might track well with the reality of testing. The fit between our simulated testing data and real testing data in our chosen case study suggests the model is capturing the true epidemic peak - and also of reflecting the impact of a testing regime. Exploration of different testing regimes represents a promising future direction for research. Regardless, researchers should ensure that modelled results distinguish between cases and *reported* cases, and should seek to document the statistical process which mediates the relationship between these. Reported case numbers will paint only a partial picture of the full situation, but through simulation we may begin to better understand the underlying reality.

References

- 1 National Statistics Agency. Zimbabwe Population Census 2012. https://international.ipums.org/international-action/sample_details/country/zw, 2012.
- 2 Adam T Biggs and Lanny F Littlejohn. Revisiting the initial COVID-19 pandemic projections. *The Lancet Microbe*, 2(3):e91–e92, March 2021. doi:10.1016/S2666-5247(21)00029-X.
- 3 BSG. Oxford University Government Response Tracker. <https://www.bsg.ox.ac.uk/research/covid-19-government-response-tracker>, 2020.
- 4 Sheryl L. Chang, Nathan Harding, Cameron Zachreson, Oliver M. Cliff, and Mikhail Prokopenko. Modelling transmission and control of the COVID-19 pandemic in Australia. *Nature Communications*, 11(1):5710, November 2020. doi:10.1038/s41467-020-19393-6.

- 5 Itai Chitungo, Tafadzwa Dzinamarira, Nigel Tungwarara, Munashe Chimene, Solomon Mukwenha, Edward Kunonga, Godfrey Musuka, and Grant Murewanhema. COVID-19 Response in Zimbabwe: The Need for a Paradigm Shift? *COVID*, 2(7):895–906, June 2022. doi:10.3390/covid2070065.
- 6 Tafadzwa Dzinamarira, Mathias Dzobo, and Itai Chitungo. COVID-19: A perspective on Africa’s capacity and response. *Journal of Medical Virology*, 92(11):2465–2472, November 2020. doi:10.1002/jmv.26159.
- 7 Tafadzwa Dzinamarira, Munyaradzi P. Mapingure, Gallican N. Rwibasira, Solomon Mukwenha, and Godfrey Musuka. COVID-19: Comparison of the Response in Rwanda, South Africa and Zimbabwe. *MEDICC Review*, July 2021. doi:10.37757/MR2021.V23.N3.4.
- 8 Tafadzwa Dzinamarira, Solomon Mukwenha, Rouzeh Eghtessadi, Diego F Cuadros, Gibson Mhlanga, and Godfrey Musuka. Coronavirus Disease 2019 (COVID-19) Response in Zimbabwe: A Call for Urgent Scale-up of Testing to meet National Capacity. *Clinical Infectious Diseases*, 72(10):e667–e674, May 2021. doi:10.1093/cid/ciaa1301.
- 9 Federal Research Centre for Cultivated Plants. Official Ports of Entry for Zimbabwe.
- 10 N Ferguson, D Laydon, G Nedjati Gilani, N Imai, K Ainslie, M Baguelin, S Bhatia, A Boonyasiri, ZULMA Cucunuba Perez, G Cuomo-Dannenburg, A Dighe, I Dorigatti, H Fu, K Gaythorpe, W Green, A Hamlet, W Hinsley, L Okell, S Van Elsland, H Thompson, R Verity, E Volz, H Wang, Y Wang, P Walker, P Winskill, C Whittaker, C Donnelly, S Riley, and A Ghani. Report 9: Impact of non-pharmaceutical interventions (NPIs) to reduce COVID19 mortality and healthcare demand. Technical report, Imperial College London, March 2020. doi:10.25561/77482.
- 11 Arun Fryatt, Victoria Simms, Tsitsi Bandason, Nicol Redzo, Ioana D. Olaru, Chiratidzo E Ndhlovu, Hilda Mujuru, Simbarashe Rusakaniko, Michael Hoelscher, Raquel Rubio-Acero, Ivana Paunovic, Andreas Wieser, Prosper Chonzi, Kudzai Masunda, Rashida A Ferrand, and Katharina Kranzer. Community SARS-CoV-2 seroprevalence before and after the second wave of SARS-CoV-2 infection in Harare, Zimbabwe. *EClinicalMedicine*, 41:101172, November 2021. doi:10.1016/j.eclinm.2021.101172.
- 12 FT. Financial Times Covid Tracker. <https://www.ft.com/content/a2901ce8-5eb7-4633-b89c-cbdf5b386938>, 2020.
- 13 Muchaneta Gudza-Mugabe, Kenny Sithole, Lucia Sisya, Sibongile Zimuto, Lincoln S. Charimari, Anderson Chimusoro, Raiva Simbi, and Alex Gasasira. Zimbabwe’s emergency response to COVID-19: Enhancing access and accelerating COVID-19 testing as the first line of defense against the COVID-19 pandemic. *Frontiers in Public Health*, 10:871567, July 2022. doi:10.3389/fpubh.2022.871567.
- 14 Nicolas Hoertel, Martin Blachier, Carlos Blanco, Mark Olfson, Marc Massetti, Marina Sánchez Rico, Frédéric Limosin, and Henri Leleu. A stochastic agent-based model of the SARS-CoV-2 epidemic in France. *Nature Medicine*, 26(9):1417–1421, September 2020. doi:10.1038/s41591-020-1001-6.
- 15 JHU. John Hopkins Coronavirus Resource Center. <https://coronavirus.jhu.edu/map.html>, 2020.
- 16 Cliff C. Kerr, Robyn M. Stuart, Dina Mistry, Romesh G. Abeysuriya, Katherine Rosenfeld, Gregory R. Hart, Rafael C. Núñez, Jamie A. Cohen, Prashanth Selvaraj, Brittany Hagedorn, Lauren George, Michał Jastrzębski, Amanda Izzo, Greer Fowler, Anna Palmer, Dominic Delpont, Nick Scott, Sherrie Kelly, Carrie Bennette, Bradley Wagner, Stewart Chang, As-saf P. Oron, Edward Wenger, Jasmina Panovska-Griffiths, Michael Famulare, and Daniel J. Klein. Covasim: An agent-based model of COVID-19 dynamics and interventions. Preprint, *Epidemiology*, May 2020. doi:10.1101/2020.05.10.20097469.
- 17 Imran Mahmood, Hamid Arabnejad, Diana Suleimenova, Isabel Sassoon, Alaa Marshan, Alan Serrano-Rico, Panos Louvieris, Anastasia Anagnostou, Simon J E Taylor, David Bell, and Derek Groen. FACS: A geospatial agent-based simulator for analysing COVID-19 spread and

- public health measures on local regions. *Journal of Simulation*, 16(4):355–373, July 2022. doi:10.1080/17477778.2020.1800422.
- 18 Grant Murewanhema, Trouble Burukai, Dennis Mazingi, Fabian Maunganidze, Jacob Mufunda, Davison Munodawafa, and William Pote. A descriptive study of the trends of COVID-19 in Zimbabwe from March - June 2020: Policy and strategy implications. *Pan African Medical Journal*, 37, 2020. doi:10.11604/pamj.supp.2020.37.1.25835.
 - 19 Jungsik Noh and Gaudenz Danuser. Estimation of the fraction of COVID-19 infected people in U.S. states and countries worldwide. *PLOS ONE*, 16(2):e0246772, February 2021. doi:10.1371/journal.pone.0246772.
 - 20 John Ojal, Samuel P. C. Brand, Vincent Were, Emelda A. Okiro, Ivy K. Kombe, Caroline Mburu, Rabia Aziza, Morris Ogero, Ambrose Agweyu, George M. Warimwe, Sophie Uyoga, Ifedayo M. O. Adetifa, J. Anthony G. Scott, Edward Otieno, Lynette I. Ochola-Oyier, Charles N. Agoti, Kadondi Kasera, Patrick Amoth, Mercy Mwangangi, Rashid Aman, Wangari Ng'ang'a, Benjamin Tsofa, Philip Bejon, Edwine Barasa, Matt J. Keeling, and D. James Nokes. Revealing the extent of the first wave of the COVID-19 pandemic in Kenya based on serological and PCR-test data. *Wellcome Open Research*, 6:127, February 2022. doi:10.12688/wellcomeopenres.16748.2.
 - 21 OWID. Our World in Data. <https://ourworldindata.org/coronavirus>, 2020.
 - 22 Antonio Sarría-Santamera, Nurlan Abdukadyrov, Natalya Glushkova, David Russell Peck, Paolo Colet, Alua Yeskendir, Angel Asúnsolo, and Miguel A. Ortega. Towards an Accurate Estimation of COVID-19 Cases in Kazakhstan: Back-Casting and Capture–Recapture Approaches. *Medicina*, 58(2):253, February 2022. doi:10.3390/medicina58020253.
 - 23 WHO. World Health Organization's Coronavirus Tracker. <https://covid19.who.int/>, 2020.
 - 24 Sarah Wise, Sveta Milusheva, Sophie Ayling, and Robert Manning Smith. Scale matters: Variations in spatial and temporal patterns of epidemic outbreaks in agent-based models. *Journal of Computational Science*, 69:101999, May 2023. doi:10.1016/j.jocs.2023.101999.

Geographic Analysis of Trade-Offs Between Amenity and Supply Effects in New Office Buildings

Kazushi Matsuo¹ ✉ 

University of Tsukuba, Japan

Morito Tsutsumi ✉

University of Tsukuba, Japan

Toyokazu Imazeki ✉ 

Commercial Property Research Institute, Inc., Tokyo, Japan

Abstract

The supply of new office buildings in the neighborhood both positively and negatively affects rents. This study attempts to deepen the quantitative knowledge of this trade-off relationship and estimate the correlation between new supply and rent within a specific geographic area based on a hedonic model. Although the results exhibit biases, they indicate that supply effects become apparent after construction is completed, and that they vary geographically and are related to local market characteristics.

2012 ACM Subject Classification Social and professional topics → Geographic characteristics; Applied computing → Economics; Information systems → Geographic information systems; General and reference → Empirical studies

Keywords and phrases Office rent, new office building, amenity effect, supply effect

Digital Object Identifier 10.4230/LIPICs.GIScience.2023.51

Category Short Paper

Funding This research was supported by the Sanko Office Foundation, the Obayashi Foundation and JSPS KAKENHI Grant Number 23KJ0242.

1 Introduction

With urban areas' development, numerous new office buildings have been built in the city centers. However, the impact of the supply of these new office buildings on such regions is unknown. Generally, the impact of real estate supply on a region can be explained by the trade-off between supply and amenity effects [6]. The supply effect relates to the availability of new real estate that absorbs demand and eases the upward pressure on rents. Accordingly, the filtering process used as a supply mechanism for affordable housing can cause rent to fall and result in a cascading transfer to higher-quality properties [9]. The amenity effect is related to the supply of new real estate that attracts high-income households and new amenities, thus increasing rent in that area. Particularly, the redevelopment in low-income neighborhoods can lead to gentrification, driving existing residents outside the area [3].

The trade-off between amenity and supply effects have been studied in recent years for the housing market. [4] and [1] demonstrated that the overall supply effect is stronger in the U.S. housing market. Do these results hold true for the office segment?

¹ Corresponding author



Numerous researchers, who have analyzed office market dynamics at the city or country levels, have reported that the supply effect is consistently strong in the long term. Simultaneously, in the short run, the new supply has been observed to increase and decrease rents [8]. However, this issue – at the micro-level – has been underdiscussed.

Using data from 2000 to 2022, we answer the following three questions regarding the new supply of office buildings for the Tokyo office market, which has a high concentration of office buildings worldwide:

RQ1: What is the geographical extent of the impact of new office buildings?

RQ2: How do trade-offs between supply and amenity effects vary over time?

RQ3: Do these trade-offs vary geographically?

2 Data

This study focused on the rental office market in Tokyo's 23 wards. Sanko Estate Co. Ltd. provided the data for the analysis. This included quarterly attribute data for all rental office buildings identified by Sanko Estate Co., Ltd. The data also include information on asking rent for the advertised properties. The sample size, including asking rent and excluding missing data, was 523,566.

Tokyo's 23 wards have the world's most concentrated business cities in terms of office space, with approximately 5 million tsubo (≈ 16 million m^2) of new rental office space available between 2000 and 2022, leaving approximately 13 million tsubo (≈ 43 million m^2) rentable floor space at the end of 2022.

The indicator for neighborhood new office building supply ($NN S_{it}^r$) is the ratio of the rentable floor space of new office buildings to the rentable floor space within a radius of r meters, centered on office building i at time t . Here, r is the threshold of interest representing the spatial range affected by the new supply. Considering that r is an unknown threshold, it is empirically determined using the following method:

3 Method

3.1 Variable selection

We adopted a hedonic approach to estimate the impact of the new supply. This approach was proposed by Rosen [10] and has been widely used to explore the determinants of real estate prices (rents) [11].

$$\ln R_{it} = \beta_0 + \sum_{k=1}^K X_{itk} \beta_k + NN S_{it}^r \beta_{NNS} + \varepsilon_{it} \quad (1)$$

where $\ln R_{it}$ represents the logarithmic asking rent; X_{itk} is the k th explanatory variable; ε_{it} is the error term; and $\beta_0, \beta_k, \beta_{NNS}$ are parameters. Here, β_{NNS} is the parameter of most interest, with $\beta_{NNS} > 0$ implying a strong amenity effect and $\beta_{NNS} < 0$ implying a strong supply effect. The spatial range threshold r in $NN S_{it}^r$ was determined to be from 100 to 1500 meters, based on the Akaike information criterion (AIC) minimization. See Table 1 for the details and basic statistics on the explanatory variables X_{itk} .

To answer the RQ2 question, we extended the base model. Here, we added the lagged variables of $NN S_{it}^r$ to the model from five years ago (20 quarters) to three years later (12 quarters).

$$\ln R_{it} = \beta_0 + \sum_{k=1}^K X_{itk} \beta_k + \sum_{p=-12}^{20} NN S_{i,t-p}^r \beta_{NNS,p} + \varepsilon_{it} \quad (2)$$

■ **Table 1** Variables and description.

Variable	Content	Unit	Mean	SD
Asking rent	Monthly asking rent including common area maintenance charge	yen/tsubo (log)	9.582	0.346
Area per floor	The maximum leasable area on a standard office floor (3rd floor or higher) for each building	tsubo (log)	3.833	0.904
Age	Number of years since construction	year	25.422	11.681
Stories	Number of stories above ground	floor (log)	2.020	0.347
Time to the nearest station	Time to walk to the building from the nearest station	min	3.583	2.306
Neighborhood rentable area	Rentable gross floor area in the neighborhood	tsubo (log)	11.434	1.231
Vacancy rate	Vacancy rate of neighborhood office buildings	%	0.061	0.041
Air-conditioning	=1 if a building have air-conditioning system	{0, 1}	0.976	
Seismic performance	=1 if a building have seismic performance	{0, 1}	0.018	
Structural dummy	A set of dummy variables for building structure			
Time dummy	A set of dummy variables representing the quarter of tenant recruitment			
Area dummy	A set of dummy variables for submarkets as defined by Sanko Estate [2].			

For larger office buildings, leasing activity begins before construction is completed. In such cases, supply effects may become apparent even before construction is completed. Similarly, rent increases may be associated with expectations of future regional revitalization.

Additionally, these trade-offs may vary from region to region (RQ3). Moreover, other determinants may have a less linear relationship with rent and vary spatially or non-spatially. To consider these relationships, the linear hedonic model was extended to spatially and non-spatially varying coefficient (SNVC) models [7].

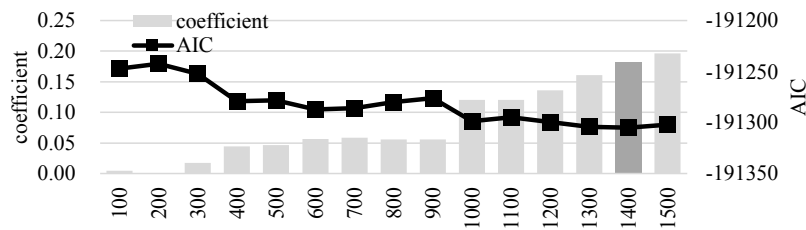
$$\ln R_{it} = f_{MC,0}(\mathbf{s}_i) + \sum_{k=1}^{K+1} X_{itk} \beta_{ik} + \varepsilon_{it}, \quad \beta_{i,k} = b_k + f_{MC,k}(\mathbf{s}_i) + g_k(X_{itk}) \quad (3)$$

where β_{ik} represents the regression coefficient and comprises the constant mean b_k , spatially varying component $f_{MC,k}(\mathbf{s}_i)$, and non-spatially varying component $g_k(X_{itk})$. The spatially varying component is a function estimated based on Moran eigenvectors, and varies with the location of property $i(\mathbf{s}_i)$. The non-spatially varying component is represented by a function that varies with the value of the variable captured by the spline function. In the SNVC model, the coefficients of each variable are selected from the constant, SVC (Spatially Varying Coefficient), NVC (Non-spatially Varying Coefficient), or SNVC, given the Bayesian information criterion (BIC) minimization. Additionally, the explanatory variable X_{itk} includes $NN S_{it}^r$. Therefore, the number of variables increases from K to $K + 1$.

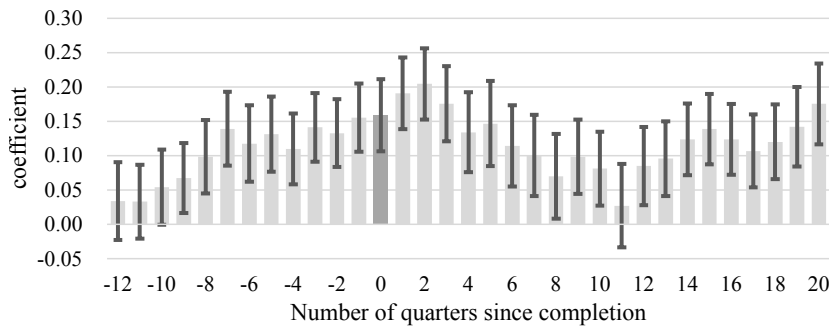
4 Result

4.1 Geographic range of impact of new supply

This section identifies the geographic range of the new supply's impact. Figure 1 depicts the change in the AIC of the model and the coefficient of the new supply when using each threshold value. The AIC is at a minimum when the radius threshold is 1400 m. Furthermore, the coefficient of the new supply is positive in all cases, and the larger the radius, the larger the absolute value of the coefficient. This suggests that the amenity effect is significant in a tradeoff relationship. However, this result is also attributable to the fact that the larger the geographic area, the smaller the percentage of new supply ($NN S_{it}^r$).



■ **Figure 1** Identification of the geographic range of the impact of neighborhood new supply on rent. The horizontal axis represents the radius (spatial range).



■ **Figure 2** Impact of neighborhood new supply before and after completion of construction.

4.2 Change in trade-offs over time

The event study graph (Figure 2) based on Equation 2 depicts the impact of the new supply in pushing rentals up for approximately three years (nine quarters). Specifically, an increase in new office stock by 10% will increase rentals by 0.7% approximately two years before construction is completed and by 1.6% upon completion. After construction is completed, the effect of rising rents declines for one to three years only to increase again. The temporary decline in impact is thought to manifest a supply effect, as tenant relocations associated with the completion of construction generate secondary vacancies. Once secondary vacancies settle, the amenity effect occurs, raising rentals to sustainable levels even after five years.

However, this interpretation requires the consideration of any remaining biases. Property developers may know the optimal locations and times to reap development profits [1]. Moreover, the new building might be planned in fast-growing areas [4]. In this case, the estimates are biased in the positive direction. The phenomenon of rents increasing two years before construction is completed is not intuitive and indicates bias. However, various actions can be taken before the new supply. Property owners may lower the rent to fill vacancies before new buildings are completed. However, if the new supply involves redevelopment, then tenants need to be temporarily relocated before construction begins. In this case, demand for office space in the neighborhood during the construction period would be temporarily increased, which might result in rising rents. While there is insufficient evidence of a strong amenity effect here, clearly, the supply effect, which becomes apparent after the construction, is weakened over time.

4.3 Spatial heterogeneity of trade-offs

Finally, the SNVC model reveals that the impact of new supply varies spatially (Figure 3). Here, the coefficient of neighborhood new supply was estimated as SVC. This result is strongly related to the aforementioned bias. Areas with significantly positive coefficients can

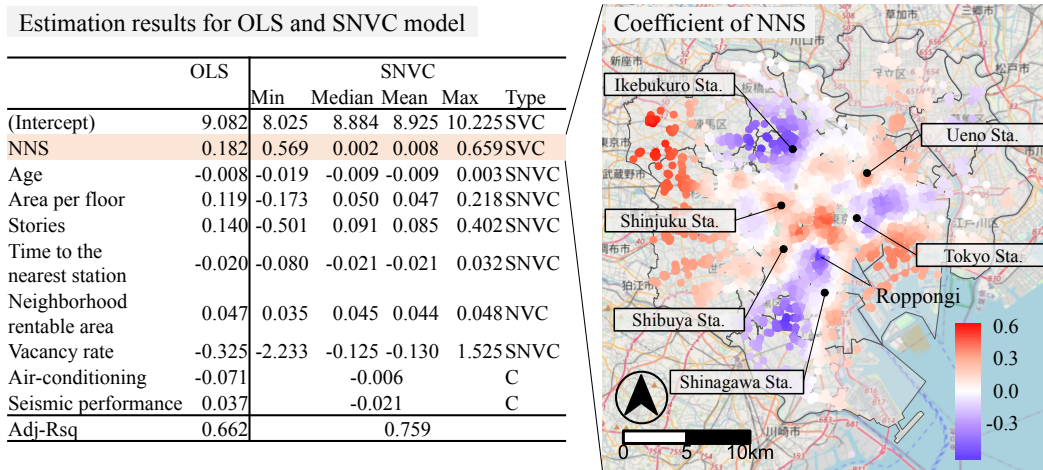


Figure 3 Estimation results and spatial distribution of NNS coefficients. Type indicates the type of coefficients, where C stands for Constant, SVC for spatially varying coefficients, NVC for non-spatially varying coefficients, and SNVC for spatially & non-spatially varying coefficients.

be interpreted as having strong amenity effect or biases. However, this does not necessarily imply that actively redeveloped areas have strong positive effects. In the case of the Roppongi and Tokyo Station areas, which underwent extensive redevelopment over the past two decades, the coefficients were either negative or zero.

This spatially heterogeneous trade-off may be related to the vacancy rate. Areas such as Shinjuku and Shibuya Sta. areas tend to have low vacancy rates in the long term, whereas Roppongi and Kanda (between Tokyo and Ueno Sta.) have high vacancy rates [5]. In localities with low vacancy rates, new buildings absorb latent demand and help boost rents, whereas in areas with high vacancy rates, secondary vacancies may become apparent and cause rents to fall.

The results of the SNVC model showed other interesting spatial heterogeneity in rent determinants, but due to volume constraints, we omit them here.

5 Conclusion

This study estimated the local impact of new office building supply. The results suggest that the model fits best when the impact of the new supply has a radius of 1400 meters. According to the results based on the linear model, the impact of the new supply was positive, but the presence of an upward bias should be considered in the discussion. However, event studies reveal that the supply effect became apparent post-construction, indicating a temporary decline in the impact of the new supply. Furthermore, the results of the SNVC model, which considers the spatial heterogeneity of the impact of the new supply, suggest that the trade-off between amenity and supply effects may be associated with high and low vacancy rates.

These results contribute to a wider discussion of the endogeneity of the new supply in terms of the location and trade-off relationship. They can be used to formulate informed policy decisions regarding office supply. If the supply effect is only temporary, the supply of quality office buildings to SMEs based on the filtering process may become complex and place financial strain on SMEs over time. However, appropriate location-based interventions are needed because of their locational variations.

Nevertheless, this study has several shortcomings as it is in its infancy. Specific strategies to remove bias and identify causation needs to be discussed. Furthermore, the new supply is interdependent on rent and vacancy rates.

References

- 1 Brian J Asquith, Evan Mast, and Davin Reed. Local effects of large new apartment buildings in low-income areas. *The Review of Economics and Statistics*, 105(2):359–375, 2023. doi:10.1162/rest_a_01055.
- 2 Sanko Estate. Office rentdata, 2023. URL: <https://www.sanko-e.co.jp/pdf/rentdata/en/market2023.pdf>.
- 3 Dan Immergluck. Large redevelopment initiatives, housing values and gentrification: The case of the atlanta beltline. *Urban Studies*, 46(8):1723–1745, 2009. doi:10.1177/0042098009105500.
- 4 Xiaodi Li. Do new housing units in your backyard raise your rents? *Journal of Economic Geography*, 22(6):1309–1352, 2022. doi:10.1093/jeg/1bab034.
- 5 Kazushi Matsuo, Morito Tsutsumi, and Toyokazu Imazeki. Spatial characteristics of the tokyo office market in terms of vacancy rates. *Theory and Applications of GIS*, 30(1):51–63, 2022 [in Japanese].
- 6 Raven Molloy. The effect of housing supply regulation on housing affordability: A review. *Regional Science and Urban Economics*, 80:103350, 2020. doi:10.1016/j.regsciurbeco.2018.03.007.
- 7 Daisuke Murakami and Daniel A Griffith. Balancing spatial and non-spatial variation in varying coefficient modeling: A remedy for spurious correlation. *Geographical Analysis*, 55(1):31–55, 2023. doi:10.1111/gean.12310.
- 8 Krzysztof Nowak, Michal Gluszak, and Stanislaw Belniak. Dynamics and asymmetric rent adjustments in the office market in warsaw. *International Journal of Strategic Property Management*, 24(6):428–440, 2020. doi:10.3846/ijspm.2020.13647.
- 9 Brendan O’Flaherty. An economic theory of homelessness and housing. *Journal of Housing Economics*, 4(1):13–49, 1995. doi:10.1006/jhec.1995.1002.
- 10 Sherwin Rosen. Hedonic prices and implicit markets: product differentiation in pure competition. *Journal of Political Economy*, 82(1):34–55, 1974. doi:10.1086/260169.
- 11 Stephen Sheppard. Hedonic analysis of housing markets. *Handbook of Regional and Urban Economics*, 3:1595–1635, 1999. doi:10.1016/S1574-0080(99)80010-8.

Impacts of Catchments Derived from Fine-Grained Mobility Data on Spatial Accessibility

Alexander Michels   

CyberGIS Center for Advanced Digital and Spatial Studies, University of Illinois Urbana-Champaign, IL, USA

Jinwoo Park 

CyberGIS Center for Advanced Digital and Spatial Studies, University of Illinois Urbana-Champaign, IL, USA

Bo Li

Department of Statistics, University of Illinois Urbana-Champaign, IL, USA

Jeon-Young Kang 

Department of Geography, Kyung Hee University, Dongdaemun-gu, Seoul, South Korea

Shaowen Wang   

CyberGIS Center for Advanced Digital and Spatial Studies, University of Illinois Urbana-Champaign, IL, USA

Abstract

Spatial accessibility is a powerful tool for understanding how access to important services and resources varies across space. While spatial accessibility methods traditionally rely on origin-destination matrices between centroids of administrative zones, recent work has examined creating polygonal catchments – areas within a travel-time threshold – from point-based fine-grained mobility data. In this paper, we investigate the difference between the convex hull and alpha shape algorithms for determining catchment areas and how this affects the results of spatial accessibility analyses. Our analysis shows that the choice of how we define a catchment produces differences in the measured accessibility which correlate with social vulnerability. These findings highlight the importance of evaluating and communicating minor methodological choices in spatial accessibility analyses.

2012 ACM Subject Classification Applied computing → Earth and atmospheric sciences; Applied computing → Health informatics; Applied computing → Transportation

Keywords and phrases Spatial accessibility, alpha shape, convex hull, cyberGIS, social vulnerability

Digital Object Identifier 10.4230/LIPIcs.GIScience.2023.52

Category Short Paper

Funding This paper is based upon work supported in part by the Institute for Geospatial Understanding through an Integrative Discovery Environment (I-GUIDE) that is funded by the National Science Foundation (NSF) under award No. 2118329. Any opinions, findings, conclusions, or recommendations expressed in this paper are those of the author(s) and do not necessarily reflect the views of NSF. The work also received support from the Taylor Geospatial Institute. Our computational experiments used Virtual ROGER that is a geospatial supercomputer supported by the CyberGIS Center for Advanced Digital and Spatial Studies and the School of Earth, Society and Environment at the University of Illinois Urbana-Champaign.

1 Introduction

Spatial accessibility is an important field of research that examines access across space to vital resources and services like healthcare [11, 12, 14]. This makes spatial accessibility a powerful tool for identifying and analyzing disparities in access across space. Access is especially



© Alexander Michels, Jinwoo Park, Bo Li, Jeon-Young Kang, and Shaowen Wang; licensed under Creative Commons License CC-BY 4.0

12th International Conference on Geographic Information Science (GIScience 2023).

Editors: Roger Beecham, Jed A. Long, Dianna Smith, Qunshan Zhao, and Sarah Wise; Article No. 52; pp. 52:1–52:6

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

crucial for socially vulnerable populations – those who are socio-economically disadvantaged, disabled, with limited transportation, etc. – who may be less likely to overcome the barriers between them and the services they need. This means that spatial accessibility work must always be cognizant of how various methodological choices impact measures of accessibility and how these different patterns of access correlate with social vulnerability.

While spatial accessibility traditionally relies on origin-destination matrices between centroids of administrative zones, recent work in spatial accessibility has created polygonal catchments from fine-grained travel data [10, 11, 14]. These works have used fine-grained point data, such as travel-time on OpenStreetMap road networks [3] and Floating Car Data (FCD) [10], to more accurately determine catchments and service areas. To calculate these catchments from point data, researchers used convex hulls in Kang et. al. [11, 12] and alpha shapes in Jiao et. al. [10]. However, convex hulls have the potential to exaggerate the catchment area as they oversimplify the shape of accessible locations.

In this paper, we examine the implications of using the convex hull and alpha shape algorithms for defining catchments in spatial accessibility analysis with a case study in Cook County, Illinois, USA. In Section 2 we discuss our methods for determining spatial accessibility and catchments and Section 3 details our data. Section 4 gives our findings for our two research questions: (1) what are the differences in the accessibility measures when we compare the two approaches and (2) how do these differences correlate with social vulnerability? Section 5 concludes with a discussion of our findings and their implications.

2 Methods

2.1 Measuring Spatial Accessibility

Spatial accessibility analyzes the distribution of supply and/or demand across space. The Enhanced Two-Step Floating Catchment Area (E2SFCA) method is a common tool for calculating spatial accessibility [13]. The first step of E2SFCA determines the weighted ratio of supply and demand (R_j) for each supply location j using Equation 1:

$$R_j = \frac{S_j}{\sum_{k \in \{d_{kj} \in D_r\}} P_k W_r} \quad (1)$$

where S_j is the degree of supply at location j , P_k is the degree of the demand or population at location k and W_r is the weight for travel-time zone r [13]. The travel-time between the supply location k and demand location j is given by d_{kj} and each step of the summation only considers supply/demand pairs k, j if the travel-time is within that step's travel-time threshold D_r ($d_{kj} \in D_r$). In the second step, each demand location sums the weighted supply-to-demand ratios of supply locations within the travel-time zones. The equation for Step 2 of the E2SFCA method is:

$$A_i = \sum_{j \in \{d_{ij} \in D_r\}} R_j W_r \quad (2)$$

where A_i is the access at demand location i and R_j, W_r are ratios and weights from step one. This yields a measure which can be interpreted as supply-to-demand ratios across space.

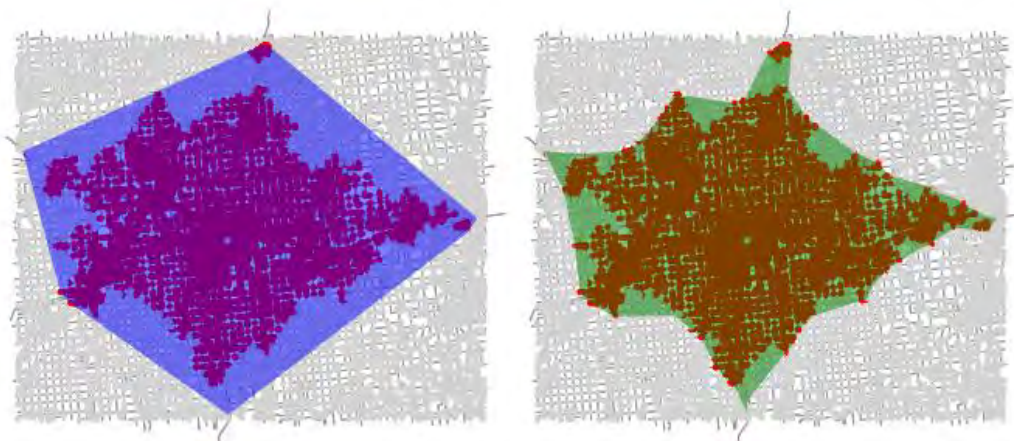
2.2 Calculating Catchments

An explosion in high-quality geospatial data and the development of cyberGIS for high-performance geospatial analysis [17] in recent years has led to a greater diversity in how travel-time catchments are defined. Mobility information is often given in the form of points

– nodes on road networks [12], mobile phone data [16], social media data [8], etc. – but there is some uncertainty in how we determine a service area from a set of points.

Our study examines two well-known approaches: convex hull [5] and alpha shapes [6]. The convex hull $\mathcal{CH}(S)$ of a set of points S is convex – meaning that the line between any two points in $\mathcal{CH}(S)$ is completely contained in $\mathcal{CH}(S)$ – and is the smallest convex set containing S [5]. Kang et. al. [12, 11] created driving-time polygons by calculating the ego-centric graph – the network around a node up to some distance threshold – on the road network around each supply location and used the convex hull to produce polygons. A similar approach was employed by Park & Goldberg using travel speed data in addition to the street network data [14]. The convex hull on a road network is given on the left of Figure 1.

Alpha shapes instead use the Delaunay triangulation of the points [6]. Using the triangulation, the alpha shape algorithm filters out triangles based on their circumradius using an alpha parameter [1]. We follow the convention used by the Python `alphashape` package [2], by filtering out triangles in the Delaunay triangulation which have a circumradius greater than $1/\alpha$. The convex hull and alpha shape are related in that the convex hull can be thought of as an alpha shape with $\alpha = 0$; the Delaunay triangulation with all triangles [6]. Whereas the convex hull is like a rubber band around the points, the alpha shape is like shrink wrap being fitted to the points, with α telling us how long to apply the heat. Jiao et. al. (2020) [9] calculated hospital service areas (HSAs) using alpha shapes and isolated forest algorithm on taxi trajectory data and Jiao et. al. (2022) [10] calculated accessibility using these service areas. An alpha shape on a road network is given on the right of Figure 1.



■ **Figure 1** An example of the difference between convex hulls and alpha shapes on road network data. The street network nodes in red are within 30 minutes of the Carle Hospital in Urbana, IL while the grey nodes are not. The convex hull around the red nodes is on the left and the alpha shape (with $\alpha = 2^{-13}$ using Albers Equal Area Conic projection) is given on the right.

To determine catchments in our experiments, we calculated travel-time with the `osmnx` package [3]. First, we cleaned the road networks to remove all but the largest weakly and strongly connected components of the network, ensuring each hospital and census tract were reachable, and determined free-flow travel-times for each edge. Distance between nodes on the graph were calculated with the Python `networkx` package using Dijkstra’s Algorithm. We collected the coordinates of each node within the travel-time threshold and created collections of points using the Python `geopandas` package. Using the collection of points, we were able to calculate polygonal catchments using the convex hull and alpha shape algorithms.

3 Study Area and Data

This study examined spatial accessibility for the general population to Intensive Care Unit (ICU) beds and its relationship with social vulnerability in Cook County, Illinois. This analysis required several different sets of data: (1) population and social vulnerability, (2) hospitals and ICU beds, and (3) road network data. Our population and social vulnerability data comes from the Centers for Disease Control (CDC) Social Vulnerability Index (SVI) which includes population estimates at the census tract level from the American Community Survey 5-Year (2014-2018) [4]. The hospitals and ICU beds per hospital were obtained from the Homeland Infrastructure Foundation-Level Data Geoplatform¹. Our road network dataset was obtained from OpenStreetMap using the Python `osmnx` package [3].

4 Results

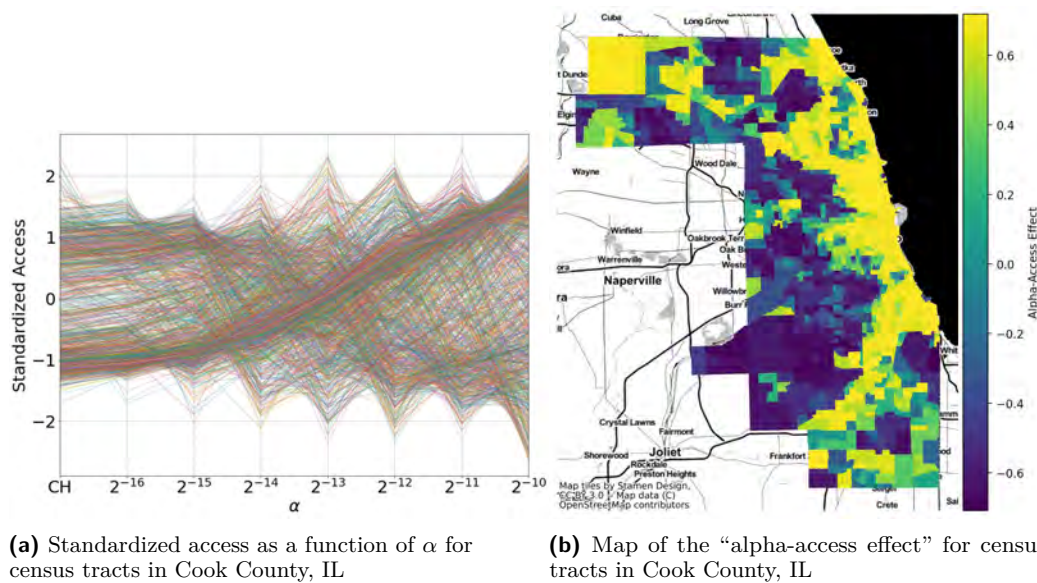
Our experiments answer two key research questions. First, what is the relationship between the alpha parameter and measured access across space? Second, does the relationship between alpha and access correlate with the CDC SVI? To accomplish this, we calculated alpha shapes using a range of alphas ($2^{-16}, 2^{-15}, \dots, 2^{-6}$) and convex hulls around the 10, 20, and 30 minute travel zones for each hospital. Then, we calculated spatial accessibility using the E2SFCA method with the catchments produced.

To understand the relationship between alpha and access, we plotted for each census tracts standardized accessibility score based on each tested α value as shown in Figure 2a. For each census tract, we compiled the distribution of spatial accessibility measures for each value of alpha and standardized the data such that the mean was zero and standard deviation is one. It is hard to determine a clear relationship here: access in some census tracts increase while others decrease as alpha rises. However, it is clear from Figure 2a that our choice of convex hull versus alpha shape makes a significant impact on the measured spatial accessibility.

To quantify the relationship between alpha and access, we computed an “alpha-access effect” metric for each census tract, mapped in Figure 2b. The metric is the linear regression coefficient between log base two of the alpha values and standardized mean accessibility for each census tract. A positive alpha-access value indicates a positive relationship between the alpha value and the measured accessibility, whereas a negative value means the measured accessibility tends to decrease as the alpha value increases. The clustering in the map prompted us to check for spatial autocorrelation and we found a Moran’s I of 0.465 and p-value of zero given by `pysal` using a Gaussian weight matrix. The map in Figure 2b shows positive values in downtown Chicago, against Lake Michigan, and generally declining values as we move west with some exceptions like O’Hare International Airport (north-western corner), the I-90 corridor (the yellow strip running from downtown Chicago to O’Hare), and the I-57 corridor (the yellow strip running south-west from the lake).

Lastly, we found a weak negative correlation (Kendall’s τ : -5.45e-02, p-value: 3.06e-03) between the alpha-access effect and SVI. This is a statistically significant result at the 0.01 significance level and indicates that census tracts with high social vulnerability tend to also be the ones where measured spatial accessibility decreases as a function of alpha. Practically, this means convex hulls and low α alpha shapes tend to over-report access in socially vulnerable communities relative to higher α alpha shapes.

¹ <https://hifld-geoplatform.opendata.arcgis.com/datasets/geoplatform::hospitals-1>



■ **Figure 2** (Left) Plots of standardized accessibility by census tracts as a function of alpha for Cook County, IL. CH stands for Convex Hull and is an alpha shape with alpha equal to zero. (Right) A map of Cook County, IL giving the alpha-access effect of each census tract.

5 Concluding Discussion

In this paper, we explored how different catchment constructions (i.e., convex hull and alpha shape) affect spatial accessibility metrics that employ granular travel-time data. Our work shows that the differences are spatially autocorrelated and vary greatly depending on the alpha value used. In addition, we demonstrated that these differences in spatial accessibility – arising from the choice between convex hulls and alpha shapes – correlate with social vulnerability. This suggests that using convex hulls and low α alpha shapes for spatial accessibility may overestimate access for socially vulnerable populations which could have unintended policy implications.

While we cannot claim that either convex hulls or alpha shapes provide a ground truth for mobility, alpha shapes with appropriate values of alpha more accurately represent the data we have, as seen in Figure 1. Therefore, we can conclude that using convex hulls and inappropriately low α alpha shapes for constructing catchments tend to over-report access to ICU beds for those who are socially vulnerable in Cook County, IL. This may lead to policy-makers providing less support to socially vulnerable populations.

There is future work to do in this vein of research as more diverse spatial datasets and tools become available. It would be illuminating to apply this methodology to a variety of cities to see how much of our findings hold in cities generally. Additionally, our work used OpenStreetMap data, but there is a variety of mobility and transportation data which have potential for use in spatial accessibility studies including the Floating Car Data [10] and temporally dynamic mobility data [15]. Further work could also help to identify the circumstances in which convex hulls and alpha shapes more accurately describe real-world mobility which varies heavily based on individual-level characteristics [7].

References

- 1 Nataraj Akkiraju, Herbert Edelsbrunner, Michael Facello, Ping Fu, EP Mucke, and Carlos Varela. Alpha shapes: definition and software. In *Proceedings of the 1st international computational geometry software workshop*, volume 63, 1995.
- 2 Ken Bellock, Neil Godber, and Philip Kahn. bellockk/alphashape: v1.3.1 Release, April 2021.
- 3 Geoff Boeing. OSMnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks. *Computers, Environment and Urban Systems*, 65:126–139, 2017.
- 4 Centers for Disease Control and Prevention/ Agency for Toxic Substances and Disease Registry/ Geospatial Research, Analysis, and Services Program. CDC/ATSDR Social Vulnerability Index 2018 Database US, 2018.
- 5 Mark de Berg, Otfried Cheong, Marc van Kreveld, and Mark Overmars, editors. *Computational Geometry: Algorithms and Applications*. Springer, Berlin, Heidelberg, 2008.
- 6 H. Edelsbrunner, D. Kirkpatrick, and R. Seidel. On the shape of a set of points in the plane. *IEEE Transactions on Information Theory*, 29(4):551–559, July 1983.
- 7 Amin Gharebaghi and Mir Abolfazl Mostafavi. Space-Time Representation of Accessible Areas for Wheelchair Users in Urban Areas (Short Paper). In Stephan Winter, Amy Griffin, and Monika Sester, editors, *10th International Conference on Geographic Information Science (GIScience 2018)*, volume 114 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 28:1–28:6, Dagstuhl, Germany, 2018. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.
- 8 Yingjie Hu and Jimin Wang. How Do People Describe Locations During a Natural Disaster: An Analysis of Tweets from Hurricane Harvey. In *11th International Conference on Geographic Information Science (GIScience 2021) - Part I*, volume 177 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 6:1–6:16, Dagstuhl, Germany, 2020.
- 9 W. Jiao, H. Fan, and Y. Wang. Analyzing the Spatiotemporal Patterns of Emergency Medical Travels from FCD Data. In *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, volume V-3-2020, pages 751–756. Copernicus GmbH, August 2020. doi:10.5194/isprs-annals-V-3-2020-751-2020.
- 10 Wei Jiao, Wei Huang, and Hongchao Fan. Evaluating spatial accessibility to healthcare services from the lens of emergency hospital visits based on floating car data. *International Journal of Digital Earth*, 15(1):108–133, December 2022. doi:10.1080/17538947.2021.2014578.
- 11 Jeon-Young Kang, Bitu Fayaz Farkhad, Man-pui Sally Chan, Alexander Michels, Dolores Albarracin, and Shaowen Wang. Spatial accessibility to HIV testing, treatment, and prevention services in Illinois and Chicago, USA. *PLOS ONE*, 17(7):e0270404, July 2022.
- 12 Jeon-Young Kang, Alexander C Michels, Fangzheng Lyu, Shaohua Wang, Nelson Agbodo, Vincent L Freeman, and Shaowen Wang. Rapidly Measuring Spatial Accessibility of COVID-19 Healthcare Resources: A Case Study of Illinois, USA. *International Journal of Health Geographics*, 2020. doi:10.1186/s12942-020-00229-x.
- 13 Wei Luo and Yi Qi. An enhanced two-step floating catchment area (E2SFCA) method for measuring spatial accessibility to primary care physicians. *Health & Place*, 15(4):1100–1107, December 2009. doi:10.1016/j.healthplace.2009.06.002.
- 14 Jinwoo Park and Daniel W. Goldberg. An Examination of the Stochastic Distribution of Spatial Accessibility to Intensive Care Unit Beds during the COVID-19 Pandemic: A Case Study of the Greater Houston Area of Texas. *Geographical Analysis*, July 2022. doi:10.1111/gean.12340.
- 15 Jinwoo Park, Alexander Michels, Fangzheng Lyu, Su Yeon Han, and Shaowen Wang. Daily changes in spatial accessibility to ICU beds and their relationship with the case-fatality ratio of COVID-19 in the state of Texas, USA. *Applied Geography*, page 102929, March 2023.
- 16 Michael Sinclair, Qunshan Zhao, Nick Bailey, Saeed Maadi, and Jinhyun Hong. Understanding the use of greenspace before and during the COVID-19 pandemic by using mobile phone app data. In *GIScience 2021*, September 2021. doi:10.25436/E2D59P.
- 17 Shaowen Wang. A CyberGIS Framework for the Synthesis of Cyberinfrastructure, GIS, and Spatial Analysis. *Annals of the Association of American Geographers*, 100(3):535–557, June 2010. doi:10.1080/00045601003791243.

Exploring the Potential of Machine and Deep Learning Models for OpenStreetMap Data Quality Assessment and Improvement

Salim Miloudi¹ ✉ 

Spatial Reference Information Systems Department, Space Techniques Center, Oran, Algeria

Bouhadjar Meguenni ✉ 

Spatial Reference Information Systems Department, Space Techniques Center, Oran, Algeria

Abstract

The OpenStreetMap (OSM) project is a widely-used crowdsourced geographic data platform that allows users to contribute, edit, and access geographic information. However, the quality of the data in OSM is often uncertain, and assessing the quality of OSM data is crucial for ensuring its reliability and usability. Recently, the use of machine and deep learning models has shown to be promising in assessing and improving the quality of OSM data. In this paper, we explore the current state-of-the-art machine learning models for OSM data quality assessment and improvement as an attempt to discuss and classify the underlying methods into different categories depending on (1) the associated learning paradigm (supervised or unsupervised learning-based methods), (2) the usage of extrinsic or intrinsic-based metrics (i.e., assessing OSM data by comparing it against authoritative external datasets or via computing some internal quality indicators), and (3) the use of traditional or deep learning-based models for predicting and evaluating OSM features. We then identify the main trends and challenges in this field and provide recommendations for future research aiming at improving the quality of OSM data in terms of completeness, accuracy, and consistency.

2012 ACM Subject Classification Information systems → Geographic information systems

Keywords and phrases OpenStreetMap (OSM), Volunteered Geographic Information (VGI), Machine Learning (ML), Deep Learning (DL), Quality Assessment (QA), Building Footprint Detection, Semantic Segmentation

Digital Object Identifier 10.4230/LIPIcs.GIScience.2023.53

Category Short Paper

Acknowledgements We would like to thank the anonymous reviewers for their valuable comments.

1 Introduction

The OpenStreetMap (OSM) project ² is a collaborative effort to create a free, editable map of the world. The OSM database is built and maintained by a community of volunteers who contribute data on various geographical features such as roads, buildings, and points of interest. For this purpose, there are various editors that can be used to edit OSM data, including web-based editors such as iD and Potlatch ³, and desktop editors such as JOSM ⁴ and Merkaartor ⁵. Each editor has its own set of features and tools, making them suitable for different types of mapping tasks. For example, JOSM is a powerful editor that has a wide

¹ Corresponding author

² <https://www.openstreetmap.org>

³ <https://www.systemed.net/potlatch/>

⁴ <https://josm.openstreetmap.de/>

⁵ <http://merkaartor.be/>



range of advanced features and is suitable for experienced mappers, while iD is a web-based editor that is easy to use and is suitable for beginners. Additionally, some editors like JOSM have plugins that can automate certain tasks, such as checking for errors in the data.

The importance of OSM data lies in its wide range of applications. It is essentially used in many fields such as navigation, emergency response, transportation and urban planning. It is also used to create custom maps for specific needs, such as hiking and biking maps, and is used as a base map for many other applications. Besides, OSM data can be used to create map tiles and other map products that can be used on websites and mobile applications. Despite its usefulness and its reliability, the quality of OSM data is strongly dependent on the accuracy and completeness of the contributions (edits or changesets) made by volunteers. In fact, as the OSM database continues to grow, the need for automated methods to assess and improve its data quality becomes increasingly important.

On the other hand, numerous machine and deep learning models have been applied to various different tasks in the area of GIS (geographic information systems) and web-mapping, including map digitization using features generated by artificial intelligence (AI) predictions. For instance, one could extract building footprint binary masks from drone imagery via different deep learning segmentation models, transform those masks into georeferenced polygons, and then overlay those geometries on OSM base-map for quality assessment. This will allow us to build AI tools capable to assist the mappers detect incomplete regions and vandalism cases when there is a mismatching between the predicted features and the existing annotations created by the contributors within a certain area of interest (AOI) in OSM.

In this paper, we review some of the state-of-the-art machine learning models for OSM data quality assessment while describing the proposed approach and the important findings for each work.

2 Machine Learning Models for OSM Data Quality Assessment

Mapping systems are crucial for navigation, transportation and other applications, but they can be costly to maintain due to the need for regular updates. Traditional maps, also known as authoritative maps, may not be updated as frequently due to budget constraints and may result in inaccuracies in terms of temporal, spatial and completeness. An alternative solution is Volunteered Geographic Information (VGI) [9], which relies on the contributions of individuals to create and update maps. One of the most popular VGI projects is OpenStreetMap (OSM), which was launched in 2004 and currently has over 10 million users from around the world. OpenStreetMap (OSM) is a VGI project which serves as an alternative to traditional map sources and is open to the public for retrieving, adding and editing spatial features. While OSM data is constantly being improved, the completeness and quality of the data may vary depending on the number of contributors and their mapping skills [19]. For instance, OSM coverage is more or less complete in urban areas compared to rural areas [10]. Additionally, it is not uncommon to encounter missing roads and inaccuracies in terms of positional accuracy [28] and semantic tags [6, 14].

Despite these limitations (i.e., issues related to data completeness and its quality), OSM has been widely used in a variety of applications, including land cover mapping and classification [5, 25, 3, 26], navigation (e.g., traffic estimation) [16], 3D city modeling and location-based services [22], building footprint detection using aerial imagery [27, 18], location-based map services [30] and indoor mapping [8].

To evaluate and improve the quality of OSM data, researchers have proposed various methods to tackle issues related to completeness [15], positional accuracy [4], semantic tag accuracy [7] and topological consistency [21]. Other works approached OSM data quality assessment [24, 13] by performing OSM meta-analysis, such as examining the activities of the contributors [20, 2].

In recent years, there has been an increasing interest in automating tasks related to OSM data. In fact, numerous works have used machine learning and remote sensing techniques to improve OSM data, while deep learning [29] has been used to extract information from OSM data to train image recognition models. Overall, the combination of machine learning, earth observation and OSM data has the potential to address global challenges in new ways. Several supervised machine learning based models have been trained on properties of OSM objects to find potential annotation errors. The authors in [1] have proposed three different machine learning based approaches to identify errors (inconsistent tags) in OSM object annotations. The first approach, *consistency checking*, involves applying a classifier while the user is editing and assigning tags to OSM objects. In this case, the editing tool can inform the volunteer if the assigned tag value is inconsistent with what the classifier predicted. Usually, geometrical, topological, and contextual properties (e.g., the object area) are used to train the supervised learning classifier. The second approach, *manual checking*, consists of applying a supervised classifier on a selected set of objects from OSM and then having OSM users to manually validate the objects whose tags present inconsistencies with the predictions of the classifier. The third approach, *automatic checking*, involves using a classifier to automatically correct annotations based on its predictions without human verification.

Conventional methods typically compare Volunteered Geographic Information (VGI) against an *authoritative* dataset to evaluate the quality of VGI data such as OSM data. While authoritative data is generated by official organizations, VGI is contributed voluntarily by individuals or communities. Also, VGI can be less reliable and accurate due to varying quality and expertise, while authoritative data is more trusted. In addition, VGI is more dynamic but lacks consistent quality control, can have biases, legal concerns, and sustainability challenges. Despite the previous limitations, combining VGI with authoritative or reference data is recommended to tackle the aforementioned shortcomings.

In cases where reference data is unavailable to assess the quality of OSM data, intrinsic methods that evaluate the data itself and its metadata can be employed. The study described in [17] utilizes unsupervised machine learning (k-means clustering algorithm) to analyze OSM history data in Mozambique, aiming to gain insights into the contributors, their timing, and their contributions. The results obtained from the analysis showed that a majority of the data in Mozambique (93%) was contributed by a small group of active contributors (25%). The study also identified a new category of contributors who were newcomers to the area, likely attracted by HOT mapping events during disaster relief operations in Mozambique in 2019. While intrinsic methods cannot serve as a substitute for ground truthing or extrinsic methods, they offer alternative means of gaining insights into data quality and can contribute to efforts aimed at enhancing it.

The study presented in [12] takes a similar approach by examining the OSM database in Ottawa-Gatineau. The focus of the investigation is on historical map features and contributor data to understand how accurately users contribute to the OSM database. To classify the changesets and OSM contributors, two unsupervised machine learning models, namely K-means and Principal Component Analysis (PCA), are utilized. The findings reveal a cluster of skilled contributors identified as OSM experts, based on their strong contribution loadings related to the use of advanced OSM editors, and weaker loadings associated with

feature creation and frequency of contributions resulting in further correction. Therefore, attributing data quality is done by identifying experienced contributors who are likely to make further corrections and improvements to the OSM database.

On the other hand, the authors in [23] introduced a deep learning approach to address the challenge of detecting buildings in areas with limited data. They achieved this by transferring a pre-trained building detection model on a well-mapped region in OSM to data-scarce areas. The transfer was accomplished through fine-tuning the model using a combination of training samples from the original and target areas. The effectiveness of the method was validated by applying a deep neural networks trained in Tanzania to a site in Cameroon. The fine-tuned model successfully identified numerous OSM buildings that were missing in a specific area of Cameroon. The results demonstrated a significant improvement in the f1-score, even with only 30 training examples from the target area.

Moreover, the paper in [11] presents a novel approach that combines deep learning and crowdsourcing using the MapSwipe⁶ platform. The authors devised a strategy for assigning classification tasks to either deep learning or crowdsourcing based on the confidence level of the derived binary classification results. They conducted three case studies in Guatemala, Laos, and Malawi to assess the effectiveness of their proposed workflow. The findings indicated that both crowdsourcing and deep learning surpassed existing earth observation methods and products like the Global Urban Footprint in terms of performance and accuracy.

3 Conclusions and Perspectives

OpenStreetMap (OSM) is a collaborative, open-source project that aims to create a free and editable map of the world. The data in OSM is contributed and maintained by a global community of volunteer mappers, who use various tools to edit and update the map. However, the data in OSM can be inconsistent and contain errors, which can lead to inaccuracies in the map. This paper has discussed the contribution of machine and deep learning models to the assessment of the OSM data quality.

In fact, various traditional machine learning models have been used in several studies to automatically detect errors in OSM data, such as annotation errors, topological errors, and positional errors. Technically, classifiers have been trained to automatically detect errors in new data, and to recommend tag values for new objects being added to the map. Additionally, other works have deployed these models to extract rules from OSM data to help with the disambiguation of geographical objects.

On the other hand, deep learning models have been used largely to segment high-resolution satellite imagery for roads and building footprint detection. The extracted features could be used later on to assess and enrich OSM data quality.

Working on improving the existing machine learning models in terms of providing better training data quality and designing & optimizing larger models will certainly play an important role in making OSM data a more valuable and reliable data source for various real world applications.

References




- 1 Ahmed Loai Ali, Falko Schmid, Rami Al-Salman, and Tomi Kauppinen. Ambiguity and plausibility: Managing classification quality in volunteered geographic information. In *Pro-*

⁶ <https://mapswipe.org/en/index.html>

- ceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, SIGSPATIAL '14*, pages 143–152, New York, NY, USA, 2014. Association for Computing Machinery. doi:10.1145/2666310.2666392.
- 2 Jamal Jokar Arsanjani and Eric Vaz. An assessment of a collaborative mapping approach for exploring land use patterns for several european metropolises. *International Journal of Applied Earth Observation and Geoinformation*, 2015. doi:10.1016/j.jag.2014.09.009.
 - 3 Nicolas Audebert, Bertrand Le Saux, and Sébastien Lefèvre. Joint learning from earth observation and openstreetmap data to get faster better semantic maps. *CoRR*, abs/1705.06057, 2017. arXiv:1705.06057.
 - 4 Hongchao Fan, Alexander Zipf, Qing Fu, and Pascal Neis. Quality assessment for building footprints data on openstreetmap. *International Journal of Geographical Information Science*, 2014. doi:10.1080/13658816.2013.867495.
 - 5 Cidália Costa Fonte, Lucy Bastin, Linda See, Giles M. Foody, and Flavio Lupia. Usability of vgi for validation of land cover maps. *International Journal of Geographical Information Science*, 2015. doi:10.1080/13658816.2015.1018266.
 - 6 Stefan Funke, Robin Schirrmeister, and Sabine Storandt. Automatic extrapolation of missing road network data in openstreetmap. In *Proceedings of the 2nd International Conference on Mining Urban Data - Volume 1392*, MUD'15, pages 27–35, Aachen, DEU, 2015. CEUR-WS.org.
 - 7 Jean-François Girres and Guillaume Touya. Quality assessment of the french openstreetmap dataset. *Transactions in Gis*, 2010. doi:10.1111/j.1467-9671.2010.01203.x.
 - 8 Marcus Goetz and Alexander Zipf. Using crowdsourced geodata for agent-based indoor evacuation simulations. *ISPRS international journal of geo-information*, 2012. doi:10.3390/ijgi1020186.
 - 9 Michael F. Goodchild. Citizens as sensors: the world of volunteered geography. *GeoJournal*, 2007. doi:10.1007/s10708-007-9111-y.
 - 10 Mordechai Haklay. How good is volunteered geographical information? a comparative study of openstreetmap and ordnance survey datasets:. *Environment and Planning B-planning & Design*, 2010. doi:10.1068/b35097.
 - 11 Benjamin Herfort, Hao Li, Sascha Fendrich, Sven Lautenbach, and Alexander Zipf. Mapping human settlements with higher accuracy and less volunteer efforts by combining crowdsourcing and deep learning. *Remote Sensing*, 11(15), 2019. doi:10.3390/rs11151799.
 - 12 Kent T. Jacobs and Scott W. Mitchell. Openstreetmap quality assessment using unsupervised machine learning methods. *Transactions in GIS*, 24(5):1280–1298, 2020. doi:10.1111/tgis.12680.
 - 13 Musfira Jilani, Michela Bertolotto, Pdraig Corcoran, and Amerah Alghanim. Traditional vs. machine-learning techniques for OSM quality assessment. In Cláudio Elízio Calazans Campelo, Michela Bertolotto, and Pdraig Corcoran, editors, *Volunteered Geographic Information and the Future of Geospatial Data*, pages 47–64. IGI Global, 2017. doi:10.4018/978-1-5225-2446-5.ch003.
 - 14 Musfira Jilani, Pdraig Corcoran, and Michela Bertolotto. Automated highway tag assessment of openstreetmap road networks. In *Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, SIGSPATIAL '14*, pages 449–452, New York, NY, USA, 2014. Association for Computing Machinery. doi:10.1145/2666310.2666476.
 - 15 Thomas Koukoletsos, Mordechai Haklay, and Claire Ellul. Assessing data completeness of vgi through an automated matching procedure for linear data. *Transactions in Gis*, 2012. doi:10.1111/j.1467-9671.2012.01304.x.
 - 16 Bill Y Lin, Frank F Xu, Eve Q Liao, and Kenny Q Zhu. Transfer learning for traffic speed prediction: A preliminary study. In *Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.
 - 17 Aphiwe Madubedube, Serena Coetzee, and Victoria Rautenbach. A contributor-focused intrinsic quality assessment of openstreetmap in mozambique using unsupervised machine

- learning. *ISPRS International Journal of Geo-Information*, 10(3), 2021. doi:10.3390/ijgi10030156.
- 18 Volodymyr Mnih and Geoffrey Hinton. Learning to label aerial images from noisy data. In *Proceedings of the 29th International Conference on International Conference on Machine Learning, ICML'12*, pages 203–210, Madison, WI, USA, 2012. Omnipress.
 - 19 Peter Mooney and Padraig Corcoran. The annotation process in openstreetmap. *Transactions in Gis*, 2012. doi:10.1111/j.1467-9671.2012.01306.x.
 - 20 Pascal Neis and Dennis Zielstra. Recent developments and future trends in volunteered geographic information research: The case of openstreetmap. *Future Internet*, 2014. doi:10.3390/fi6010076.
 - 21 Pascal Neis, Dennis Zielstra, and Alexander Zipf. The street network evolution of crowdsourced maps: Openstreetmap in germany 2007-2011. *Future Internet*, 2011. doi:10.3390/fi4010001.
 - 22 Martin Over, Arne Schilling, S. Neubauer, and Alexander Zipf. Generating web-based 3d city models from openstreetmap: The current situation in germany. *Computers, Environment and Urban Systems*, 2010. doi:10.1016/j.compenurbsys.2010.05.001.
 - 23 J. Pisl, H. Li, S. Lautenbach, B. Herfort, and A. Zipf. Detecting openstreetmap missing buildings by transferring pre-trained deep neural networks. *AGILE: GIScience Series*, 2:39, 2021. doi:10.5194/agile-giss-2-39-2021.
 - 24 Hansi Senaratne, Amin Mobasheri, Ahmed Loai Ali, Cristina Capineri, and Mordechay Haklay. A review of volunteered geographic information quality assessment methods. *International Journal of Geographical Information Science*, 2017. doi:10.1080/13658816.2016.1189556.
 - 25 Shivangi Srivastava, John E. Vargas Muñoz, Sylvain Lobry, and Devis Tuia. Fine-grained landuse characterization using ground-based pictures: a deep learning solution based on globally available data. *International Journal of Geographical Information Science*, 2018. doi:10.1080/13658816.2018.1542698.
 - 26 Shivangi Srivastava, John E. Vargas-Munoz, and Devis Tuia. Understanding urban landuse from the above and ground perspectives: A deep learning, multimodal solution. *Remote Sensing of Environment*, 2019. doi:10.1016/j.rse.2019.04.014.
 - 27 John E. Vargas-Munoz, Sylvain Lobry, Alexandre X. Falcão, and Devis Tuia. Correcting rural building annotations in openstreetmap using convolutional neural networks. *Isprs Journal of Photogrammetry and Remote Sensing*, 2019. doi:10.1016/j.isprsjprs.2018.11.010.
 - 28 Yongyang Xu, Zhanlong Chen, Zhong Xie, and Liang Wu. Quality assessment of building footprint data using a deep autoencoder network. *International Journal of Geographical Information Science*, 2017. doi:10.1080/13658816.2017.1341632.
 - 29 Xiao Xiang Zhu, Devis Tuia, Lichao Mou, Gui-Song Xia, Liangpei Zhang, Feng Xu, and Friedrich Fraundorfer. Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE Geoscience and Remote Sensing Magazine*, 2017. doi:10.1109/mgrs.2017.2762307.
 - 30 Alexander Zipf, Steffen Neubauer, G Walenciak, Martin Over, Pascal Neis, and Arne Schilling. Interoperable location based services for 3d cities on the web using user generated content from openstreetmap. *Urban and Regional Data Management*, 2009. doi:10.1201/9780203869352.ch7.

On the Cartographic Communication of Places

Franz-Benjamin Mocnik   

University of Twente, Enschede, The Netherlands

Paris Lodron University of Salzburg, Austria

Abstract

Maps are excellent as a medium for communicating spatial configurations at geographical scales. However, the communication of thematic qualities of geographical features is constrained by the traditionally assumed strict classification of features on the map and the strong focus on spatial representation. This is despite the fact that places are central aspects of everyday life that we use to structure our experiences and thus the need to include them in many maps. This paper explores how places can be communicated through the map medium. In particular, it addresses the question of the extent to which places are mediated or merely referenced, and the extent to which maps already communicate places through its inherent spatial and thematic aspects. This is followed by a discussion of how maps not only communicate but also shape places. In perspective, this contributes to a better and more targeted representation of places, especially through maps, but also advances our understanding of how places are conceptually entangled with spatial and thematic aspects.

2012 ACM Subject Classification –

Keywords and phrases representation, reference, mediation, intentionality, conceptualization, maps

Digital Object Identifier 10.4230/LIPIcs.GIScience.2023.54

Category Short Paper

1 Introduction

Places are fundamental to our everyday lives because they are among the basic units we use to structure geographical space. In this context, ‘home’ and ‘work place’ take on central roles, which is why they are also referred to as first and second place, respectively [25]. This structuring role is also evident in narratives, which usually engage in one or more places. Previous research has investigated how narratives can be communicated by means of maps [28, 4, 14, 34, 17], especially also in relation to places [20, 7, 6], and what systematic problems exist in this regard [23]. Notwithstanding the results achieved related to the map medium, the content-related communication of narratives and places often resorts to the text or image medium because it would be difficult to do so with traditional, cartographic means [23]. This is despite the fact that maps contain many indications that refer to places.

Before resolving this apparent contradiction, we first delineate the terms ‘place’, ‘Point of Interest (POI)’, and further ones (Section 2). A discussion of the stylistic devices to communicate places in maps and other data sets follows (Section 3). This train of thought resolves the apparent contradiction between the numerous reference to places contained in the map and the simultaneously existing problems of communicating places cartographically. Subsequently, we argue that the communication of places by means of the map medium is not at all unidirectional but also possess performative qualities (Section 4). The paper concludes with a summary of the resulting consequences for maps and places (Section 5).

2 Place, POI, and Related Terms

This section engages with the concept of place and adjacent concepts to create a basic understanding for the framework of this paper. The explanations must not, however, be regarded as definitions of these very concepts, because the latter cannot be exhaustive due to the brevity of this section and could thus only be inadequate as a definitions.



© Franz-Benjamin Mocnik;

licensed under Creative Commons License CC-BY 4.0

12th International Conference on Geographic Information Science (GIScience 2023).

Editors: Roger Beecham, Jed A. Long, Dianna Smith, Qunshan Zhao, and Sarah Wise; Article No. 54; pp. 54:1–54:6

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Absolute or geometrical space refers to the physically measurable structure of space, as is represented by the mathematical concepts of Euclidean space and manifolds, e.g., in relation to the surface of the Earth. The *geographical or socially constructed space* is distinguished from this, since it is not given by the physical properties in the sense of a container space but only constructed by the geographical features and the being lived of these. It is thus mostly conceptualized as a *relational space*. Terms of absolute space such as ‘location’ and ‘area’ make sense only to a limited extent here, because they cannot be transferred without further ado. In geography, many further concepts of space are used in addition to these [30].

The concept of place is complex. Although there is agreement about its meaningfulness and basic idea in many respects, many different characterizations exist [5, 2]. These have in common that they ascribe meaning to a place [29], which makes it experienceable as an entity in its own right. Place are often (though not always) understood in the context of everyday behaviour and routine. Some places are socially constructed and shared, while others emerge individually without social influence. Place identity [26, 27], place attachment [35, 18, 33], sense of place [35, 15], recurring patterns of behaviour (place ballets) [31, 32], and further qualities are used to characterize places. This complexity in the description of a place without reference to absolute space distinguishes the concept of place from a *Point of Interest (POI)*.

In a way, the concepts of region and of place can be considered similar in that both refer to characteristics of geographical space. A region focusses on the common characteristics of all sites within that region that make possible the demarcation from surrounding regions, while the concept of place seeks a more holistic understanding. The latter thus commonly refers to its essence, such as its identity, place attachment, and alike. In this sense, a region refers to an extended part of space, while a place exists as such in space.

3 Places Are Communicated Through Maps

Maps reference places in many ways, despite traditionally following the absolute space paradigm [24]. Among such reference are, most prominently, place names, but many further indications of such reference exist. To better understand the nature of these indications, we introduce below two dimensions to describe and categorize them, followed by examples.

When considering maps as spatial arrangements of symbols interpreted by the map reader, maps only represent and mediate places but they cannot contain them. A distinction can be made here between two prototypical cases that represent the ends of a spectrum rather than a collection of discrete categories: the *referencing* and the *mediation* of a place [22]. In the first case, what is displayed on the map affords to establish a relationship between the map content and a place without, however, going into more detail about its qualities. This relationship only refers to the place as a whole. This is in contrast to the second case, in which some of the qualities of the place are conveyed, thus enabling the map reader to gain an impression of the place even without (or with little) previous experience of it.

The referencing and the mediation of a place is demarcated by the extent to which the qualities of the place are referred to and conveyed. The mediation refers to at least some of the qualities of the place, as opposed to the referencing. These qualities of the place referred to must, in turn, be conveyed in the map through appropriate references themselves, because platial qualities can only in very few cases apply to a map or its elements themselves. The map will hardly constitute the identity of a place, nor will it evoke the same emotions without according reference, et cetera. In the end, the concepts of referencing and mediation seem not to be qualitatively different; they only represent different levels of referencing – with regard to the place itself, or, in the case of mediation, with regard to its qualities.

■ **Table 1** Typical examples of place representations in a map, categorized by the two dimensions referencing/mediation and intentionality.

	referencing	mediation
intentional	<ul style="list-style-type: none"> ▪ label with a place name ▪ icon indicating a Point of Interest (POI) 	<ul style="list-style-type: none"> ▪ signs indicating a place ballet ▪ use of colours and symbols to convey emotions
unintentional	<ul style="list-style-type: none"> ▪ representation of a geographical feature that is reminiscent of an individually constructed place 	<ul style="list-style-type: none"> ▪ social entities and locale represented ▪ structure and arrangement of features in the map

The communication of places can also be characterized by the *intentionality* to reference or mediate the place. Place name labels, like most symbols, are consciously and intentionally¹ added to a map. Yet, whether a symbol is understood as a reference to a place depends on the map reader. A building or lake depicted on a map may, e.g., remind the reader of his or her home or favourite bathing spot. Such communication of a place can be despite the original intention to communicate other types of information. Combining these two dimensions – referencing/mediation and intentionality – results in four prototypical cases (Table 1).

Place names and POIs are prominent examples of intentional reference to places. Corresponding labels and icons serve as reference to the place, but without being a place themselves. In particular, a POI can be seen as a proxy for the communication of a place, although it is itself conceptually significantly different from a place.² It is interesting to note that the labels that refer to place names as well as icons that represent POIs are point features. Despite places having spatial characteristics and their cartographic communication often being intentional [13, 36, 8, 9], these spatial characteristics are not well represented in maps apart from the indication of a location. This demonstrates that spatial characteristics are often not considered to be among the relevant key characteristics of these places represented.

The intentional mediation of places through the map medium is difficult to achieve, which is why most attempts resort to other media such as texts, photos, and videos. Only few refer to the qualities of places, such as as embodied experience (by using the human sensory system) [20], emotions [11, 3], the inner structure of a place [8], and place ballets [10].

Practically every geographical feature displayed on the map can unintentionally become reminiscent of a place. In particular, when a place is not socially constructed and thus not shared, the map maker cannot have intended to reference or even mediate the place as it is experienced exclusively by the map reader. Examples include the road bend where I stop every day to feed the ducks; and my favourite spot at a nearby hill to which I use to retreat when I want to be alone. In many cases, the qualities of a place are mediated by the composition and spatial arrangement of the associations to the geographical features represented, and the locale and the represented socially lived features define a structure

¹ Intentionality is always accompanied by conscious communication. Conversely, non-intentionality often but not necessarily means unconscious communication of the place.

² If the map maker includes, e.g., a ‘fish & chip’ shop as a POI, then he or she (intentionally) refers to the affordance of buying fish and chips at this location. Such affordance can be assumed to be of rather long-term nature, suggesting that individuals live and experience this location as a place. In this respect, it can be assumed that not only the POI but also the place is referenced to some degree.

reminiscent of relational and thus also geographical space (cf., space syntax; [12]). An example is the partially reflected public–private space dichotomy [19]. Also, the familiarity with a place can be conjectured to potentially influence the way it is represented.

4 The Map Creation–Conceptualization Creation Cycle

Given the difficulties to communicate places through the map medium, one might assume that the latter has little influence on the communicated place itself. Many use cases, however, utilize the performative qualities of maps [1] and thus their potential influence on how we conceptualize and, ultimately, shape and live places. Maps used in urban planning represent, e.g., often a not yet existing state of the urban environment that is to be planned, evaluated, and actively shaped. Besides rendered photography-like images, the spatial arrangement and type of planned features depicted in the map provide an idea of how the place *might* feel in case of later realization, its identity, et cetera. This, in turn, can influence the to-be-developed place, thus creating a feedback loop from our mental conceptualization of the place to the map and then back to the conceptualization. In this sense, maps can serve as ‘place shapers’.

Maps can generally be assumed to have much less influence on the shaping of a place if they have not specifically been created for this purpose. In extreme cases, however, a map can make entire places come into existence, as was the case with the ‘paper town’ of Agloe, NY. A corresponding point symbol with attached label was depicted as a copyright trap on a map without such a place actually existing in reality. If someone were to mistakenly reproduce Agloe when copying the map without permission, the appearance of the place name on the new map would be an indication of copyright infringement. After another map that included Agloe actually appeared, it turned out that the place indeed existed. The place name had served as the name-giver for a petrol station and a supermarket that had only been built afterwards [16].

5 Consequences and Conclusion

Maps can communicate places by referencing or even mediating them. We have argued the latter to be particularly difficult when certain qualities of the place shall intentionally be emphasized in the map. In the following, we discuss three consequences of this fact.

First, the limited ability to convey places by means of traditional maps implies that that narratives can hardly be spanned when using this medium. Places and narratives become, in turn, more relevant when employing non-traditional map paradigms [7, 6], suggesting the exploration of alternative modes of representation beyond the traditional map paradigm [23].

Secondly, maps have limited affordances to convey platial qualities, especially idiosyncratic or socially constructed ones. The reason behind is that maps are geared to absolute space and therefore afford particularly well those tasks that refer to such space. Although many of the tasks we perform with a map *seem* to primarily relate to abstract space, they often refer to socially constructed space and places. In this sense, the tasks we actually perform with a map while employing an abstract space paradigm often need to be considered simplifications of, and thus proxies for, more complex tasks. Route finding tasks in the context of sight seeing or during ones holidays, e.g., refer to places and their characteristics rather than to pure distances and directions in absolute space, because the route chosen from one sight to another should, ultimately, not lead through filthy streets or industrial areas.

Thirdly, the map maker is in a dilemma. Places impact map creation such as in terms of how geographical features are spatially and thematically represented. This influence impairs readability as it is often obscured by thematic and spatial generalization and thus rarely evident to the map reader. If, however, a multiplicity of individualistic places without strong generalization were included, this would reduce readability as well. The emphasis on spatial aspects (as opposed to individual, platial qualities) must therefore inevitably limit readability.

The three consequences discussed demonstrate limitations when it comes to the cartographic representation of places, which is despite the need for better communication of these. This is in line with the larger picture of Platial Information Systems (PISs) and Theories of Platial Information (ToPIs), which face similar problems: the difficulty to represent the thematic diversity of places (strength of a PIS) and the difficulty to enable a high complexity in the formal reasoning about places (complexity of a PIS) [21]. Maps cannot fully solve these problems but may yet play an important role through building a bridge between formal data and human cognition. If alternative map paradigms make the qualities of places accessible to human cognition, this can contribute to solving the aforementioned problems of PISs.

Beyond the outlined consequences related to the cartographic communication of places as often individually and emotionally shaped geographical features, the question arises whether the nature of the map medium in itself has an influence on our conceptualization of places. Accordingly, the Sapir–Whorf hypothesis [37], which originally stems from the theory of linguistic relativity, can (and should) be posed here with regard to the map medium: how does the structure of a map used according to the traditional map paradigm influence our conceptualization of places and ultimately also the places themselves?

References

- 1 MB Aalbers. Do maps make geography? Part 1: redlining, planned shrinkage, and the places of decline. *ACME*, 13(4):525–556, 2014.
- 2 JA Agnew. Representing space. Space, scale and culture in social science. In JS Duncan and D Ley, editors, *Place/culture/representation*, pages 251–271. Routledge, 1993.
- 3 EP Bogucka, M Constantinides, LM Aiello, D Quercia, W So, and M Bancilhon. Cartographic design of cultural maps. *IEEE Comput Graphics and Appl*, 40(6):12–20, 2020. doi:10.1109/MCG.2020.3026596.
- 4 EE Boschmann and E Cubbon. Sketch maps and qualitative GIS: using cartographies of individual spatial narratives in geographic research. *The Professional Geogr*, 66(2):236–248, 2014. doi:10.1080/00330124.2013.781490.
- 5 T Cresswell. *Place. A short introduction*. Blackwell, 2004.
- 6 C Dolma. Reclaiming place through marginalized narratives. A critical geography and humanistic approach to the cartographic visualization of Beyoğlu, Istanbul. MSc thesis, U of Twente, 2021.
- 7 C Dolma. Reclaiming place through marginalized narratives: a critical geography and humanistic approach to the cartographic visualization of Beyoğlu, Istanbul. *3rd Int Symp on Platial Inf Sci (PLATIAL'21)*, pages 35–40, 2022. doi:10.5281/zenodo.5767184.
- 8 M Glebova. Town and gown: visualising university neighbourhoods as places within the urban environment. The example of three universities in Moscow. MSc thesis, U of Twente, 2021.
- 9 M Glebova. Visualizing fuzzy boundaries of city neighbourhoods. *3rd Int Symp on Platial Inf Sci (PLATIAL'21)*, pages 41–47, 2022. doi:10.5281/zenodo.5767186.
- 10 L Harvey. Improving the cartographic visualization techniques of platial features – the example of London parks. MSc thesis, U of Twente, 2020.
- 11 E Hauthal, A Dunkel, and D Burghardt. Emojis as contextual indicants in location-based social media posts. *ISPRS Int J Geo-Inf*, 10(6), 2021. doi:10.3390/ijgi10060407.
- 12 B Hillier and J Hanson. *The social logic of space*. Cambridge University Press, 1984. doi:10.1017/CB09780511597237.

- 13 H Hobel, P Fogliaroni, and AU Frank. Deriving the geographic footprint of cognitive regions. *19th AGILE Conf on Geogr Inf Sci*, pages 67–84, 2016. doi:10.1007/978-3-319-33783-8_5.
- 14 AY Ishola. An empirical evaluation of the story focus concept – the example of a map telling the story of “the legend of Meng Jiangnu”. MSc thesis, U of Twente, 2020.
- 15 G Kyle and G Chick. The social construction of a sense of place. *Leisure Sci*, 29(3):209–225, 2007. doi:10.1080/01490400701257922.
- 16 J Lackie. Copyright trap. *NewScientist*, 192(2574):62, 2006. doi:10.1016/S0262-4079(06)60797-5.
- 17 NA Landaverde Cortés. A conceptual framework for interactive cartographic storytelling. MSc thesis, U of Twente, 2018.
- 18 SM Low and I Altman. Place attachment. A conceptual inquiry. In I Altman and SM Low, editors, *Place attachment*, pages 1–12. Plenum, 1992. doi:10.1007/978-1-4684-8753-4_1.
- 19 M Mayer, DW Heck, and FB Mocnik. Shared mental models as a psychological explanation for converging mental representations of place – the example of OpenStreetMap. *2nd Int Symp on Platial Inf Sci (PLATIAL’19)*, pages 43–50, 2020. doi:10.5281/zenodo.3628871.
- 20 K McLean. Smell map narratives of place—Paris. *New Am Notes Online*, 6, 2013.
- 21 FB Mocnik. Putting geographical information science in place – towards theories of platial information and platial information systems. *Prog in Hum Geogr*, 46(3):798–828, 2022. doi:10.1177/03091325221074023.
- 22 FB Mocnik. On the representation of places. *GeoJournal*, 2023. doi:10.1007/s10708-023-10831-8.
- 23 FB Mocnik and D Fairbairn. Maps telling stories? *The Cart J*, 55(1):36–57, 2018. doi:10.1080/00087041.2017.1304498.
- 24 FB Mocnik and L Kühn. (Un)represented places – a case study of two sports venues in Gelsenkirchen and Dortmund. *3rd Int Symp on Platial Inf Sci (PLATIAL’21)*, pages 25–30, 2022. doi:10.5281/zenodo.5767180.
- 25 R Oldenburg. *The great good place. Cafés, coffee shops, bookstores, bars, hair salons and other hangouts at the heart of a community*. Paragon, 1989.
- 26 HM Proshansky. The city and self-identity. *Environ and Behav*, 10(2):147–169, 1978. doi:10.1177/0013916578102002.
- 27 E Relph. *Place and placelessness*. Pion, 1976.
- 28 RE Roth. Cartographic design as visual storytelling: synthesis and review of map-based narratives, genres, and tropes. *The Cart J*, 58(1):83–114, 2021. doi:10.1080/00087041.2019.1633103.
- 29 M Saar and H Palang. The dimensions of place meanings. *Living Rev in Landscape Res*, 3:3, 2009. doi:10.12942/lr1r-2009-3.
- 30 M Schroer. Spatial theories/social construction of spaces. In AM Orum, editor, *The Wiley Blackwell encyclopedia of urban and regional studies*. Wiley, 2019. doi:10.1002/9781118568446.eurs0313.
- 31 D Seamon. *A geography of the lifeworld*. Croom Helm, 1979. doi:10.4324/9781315715698.
- 32 D Seamon and C Nordin. Marketplace as place ballet. A Swedish example. *Meddelanden från Göteborgs Univ Geogr Inst. Ser B*, 67:35–41, 1980.
- 33 JS Smith, editor. *Explorations in place attachment*. Routledge, 2017. doi:10.4324/9781315189611.
- 34 L Tateosian, M Glatz, and M Shukunobe. Story-telling maps generated from semantic representations of events. *Behav & Inf Technol*, 39(4):391–413, 2020. doi:10.1080/0144929X.2019.1569162.
- 35 YF Tuan. *Space and place. The perspective of experience*. University of Minnesota Press, 1977.
- 36 R Westerholt, M Gröbe, A Zipf, and D Burghardt. Towards the statistical analysis and visualization of places. *10th Int Conf on Geogr Inf Sci (GIScience)*, pages 63:1–63:7, 2018. doi:10.4230/LIPIcs.GISCIENCE.2018.63.
- 37 B Whorf. *Language, thought, and reality*. MIT Press, 1956.

Resiliency: A Consensus Data Binning Method

Arpit Narechania ✉ 

Georgia Institute of Technology, Atlanta, GA, USA

Alex Endert ✉ 

Georgia Institute of Technology, Atlanta, GA, USA

Clio Andris ✉ 

Georgia Institute of Technology, Atlanta, GA, USA

Abstract

Data binning, or data classification, involves grouping quantitative data points into bins (or classes) to represent spatial patterns and show variation in choropleth maps. There are many methods for binning data (e.g., natural breaks, quantile) that may make the same data appear very different on a map. Some of these methods may be more or less appropriate for certain types of data distributions and map purposes. Thus, when designing a map, novice users may be overwhelmed by the number of choices for binning methods and experts may find comparing results from different binning methods challenging. We present **resiliency**, a new data binning method that assigns areal units to their most agreed-upon, consensus bin as it persists across multiple chosen binning methods. We show how this “smart average” can effectively communicate spatial patterns that are agreed-upon across binning methods. We also measure the variety of bins a single areal unit can be placed in under different binning methods showing fuzziness and uncertainty on a map. We implement resiliency and other binning methods via an open-source JavaScript library, **BinGuru**.

2012 ACM Subject Classification Human-centered computing → Geographic visualization

Keywords and phrases data binning, data classification, choropleth maps, geovisualization, geographic information systems, geographic information science, cartography

Digital Object Identifier 10.4230/LIPIcs.GIScience.2023.55

Category Short Paper

Supplementary Material

Software (Source Code): <https://github.com/arpitnarechania/binguru>

Interactive Resource (Observable Notebook): <https://observablehq.com/@arpitnarechania/binguru-demo>

1 Introduction

Data binning (or classification) is the process of grouping quantitative data values into bins (or classes), that are then represented by different colors, shades, textures, or sharpness to show spatial patterns or variations in choropleth maps [4]. A classic example of a choropleth map may be a country’s states shaded light or dark according to their low or high population, respectively. To create such maps, the population values may be placed into groups such as *high*, *medium* and *low population* using a binning method that assigns each state to a group. Using one binning method, a single state may be classified as *high population*, but using another binning method, it may be classified as *low population*. This affects the reader’s interpretation of the state and how resources may be allocated to the state.

Many binning methods exist [1, 9, 3, 14] and are built into popular GIS tools and libraries [5, 15, 12, 18]. They have strengths and weaknesses that make them (un)suitable for certain types of data distributions and map purposes. For example, *standard deviation* emphasizes normality and regions of high and low deviation from the mean; *quantile* evenly distributes data values into bins irrespective of the data distribution, highlighting regional



© Arpit Narechania, Alex Endert, and Clio Andris;
licensed under Creative Commons License CC-BY 4.0

12th International Conference on Geographic Information Science (GIScience 2023).

Editors: Roger Beecham, Jed A. Long, Dianna Smith, Qunshan Zhao, and Sarah Wise; Article No. 55; pp. 55:1–55:7

Leibniz International Proceedings in Informatics



LIPIC Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

differences; *pretty breaks* rounds off bin extents, making them visually appealing and easy-to-interpret; and *natural breaks* can capture organic data groupings and reveal outliers. Choosing an appropriate binning method is important to ensure that the map effectively represents the data and communicates information to the reader [11]. However, this determination can be overwhelming for users, particularly those who are not well-versed in statistical or cartographic concepts. Even experts may find it challenging to compare and contrast results of different binning methods for data-driven decision making.

In this paper, we present **resiliency**, a new data binning method that assigns areal units to their most agreed-upon, consensus bin as it persists across multiple methods. For example, if a county is placed into bin #2 (i.e., second to lowest group of values) across a majority of binning methods, *resiliency* will attempt to place it in that bin. We also show how *resiliency* allows for spatial patterns to be communicated with(out) outliers and how it promotes reflection on the fuzziness that binning imposes during mapping. Note that *resiliency* is neither a panacea nor a prescriptive measure, but provides more insight into a dataset. It detects which areal units are likely to “hop” across bins upon switching to a different binning approach. As a consensus method, *resiliency* can help the user be more confident in choosing a familiar binning method (e.g., natural breaks) if its result resembles resiliency. We share *resiliency* and 17 other established binning methods via **BinGuru**, an open-source JavaScript library for developers to create custom geospatial applications, offering more variety than existing GIS tools¹ and libraries².

2 Related Work

While *continuous* or *unclassed* maps are valuable for maintaining exact numeric data relationships to the visual variable [19], it is more common to group areal units into bins to reveal patterns across geographic space. In terms of methods, *equal interval*, *natural breaks*, *standard deviation*, *quantiles*, and *pretty breaks* are particularly common [3]. Genetic algorithms [1] and proximity-based [9] binning methods, which promote spatially compact and homogeneous regionalization on maps, have also been explored but are not widely used in practice. OSCAR is a human-centered binning method that leverages usage information from visualization dashboards to suggest common bin sizes for an attribute [14].

Determining an appropriate binning method often depends on several factors such as the data distribution (e.g., *standard deviation* for normally distributed data), tacit and domain specific knowledge (e.g., *manual interval* or *diverging bins* centered around a certain meaningful baseline value) [16], or a desire to have the same number of data points in each group (e.g., *quantiles*). According to Brewer and Pickle [3], *quantile* and *minimum boundary error* are best suited for general reading of epidemiological rate maps followed by *natural breaks* and a hybrid version of *equal interval*. According to Smith [17], *quantile*, *equal interval*, *standard deviation*, and *natural breaks* are accurate for data sets with specific distributional characteristics, but none of them accurately bin all types of distributions. Prior work has also explored diverse approaches and measures for assessing map complexity, emphasizing their impact on cognitive load, readability, and visual effectiveness [7, 2]. For example, Monmonier [10] found that round-number bin breaks, which are easy to read and remember, can constrain the outputs of optimization algorithms that have more significant digits than the map user would prefer or that the precision of the data warrants.

¹ ArcGIS [5] (9 binning methods), QGIS [15] (6)

² ArcGIS Maps SDK [6] (6 binning methods), Python’s PySAL [12] (10), and R’s tmap [18] (9)

Algorithm 1 Resiliency.

```

1 input : data values  $\mathbf{V}$ , binning methods  $\mathbf{M}$ , binning options  $\mathbf{O}$ 
2 output: resiliency bin breaks  $\mathbf{RB}$ 
3 // Compute bin breaks for all  $\mathbf{M}$ 
4 bin breaks  $\mathbf{B} \leftarrow \{ \}$ 
5 for method  $m$  in  $\mathbf{M}$  do
6   |  $\mathbf{B}[m] = \text{COMPUTE BINS}(\mathbf{V}, \mathbf{O}, m)$ 
7 // Determine bins for all  $\mathbf{V}$  across all  $\mathbf{M}$ 
8 bin ids  $\mathbf{ID} \leftarrow \{ \}$ 
9 for value  $v$  in  $\mathbf{V}$  do
10  | for method  $m$  in  $\mathbf{M}$  do
11  |   |  $\mathbf{ID}[v][m] = \text{ASSIGN BIN}(v, \mathbf{B}[m])$ 
12 // Compute the frequency of each bin among all  $\mathbf{M}$ 
13 bin frequencies  $\mathbf{BF} \leftarrow \{ \}$ 
14 for value  $v$  in  $\mathbf{V}$  do
15  |  $\mathbf{BF}[v] = \text{COMPUTE BIN FREQUENCY}(\mathbf{ID}[v])$ 
16 // Place values in their most frequent bins
17 most frequent bins  $\mathbf{MFB} \leftarrow \{ \}$ 
18 for value  $v$  in  $\mathbf{V}$  do
19  |  $\mathbf{MFB}[v] = \text{COMPUTE MOST FREQUENT BIN}(\mathbf{BF}[v])$ 
20 // Compute Resiliency
21 resiliency bin breaks  $\mathbf{RB} \leftarrow \{ \}$ 
22 working bin assignments  $\mathbf{WFB} \leftarrow \mathbf{MFB}$ 
23 while  $\text{VALIDATE BINS}(\mathbf{RB})$  do
24  |  $\mathbf{RB}, \mathbf{WFB} = \text{RESOLVE CONFLICTS}(\mathbf{WFB}, \mathbf{RB})$ 
25 return  $\mathbf{RB}$ 

```

3 Resiliency

Given the diversity and complexity of established binning methods, we propose a new method, *resiliency*, that assigns areal units to their most agreed-upon, consensus bin across multiple methods. Algorithm 1 illustrates the pseudo code for this method.

First, we compute the bin breaks for multiple *comparable*³ binning methods (Lines 3 - 6). For each areal unit, we track the ID (or index) of the bin (*binID*) that it was assigned to across the binning methods (Lines 7 - 11). Next, we compute the frequency of the assigned *binIDs* (Lines 12 - 15), i.e., the number of times it is placed across different *binIDs*. For each areal unit, we then compute the frequency of the most frequently assigned *binIDs* (Lines 16 - 19). The output is the *resiliency bin count* (number of bins), *interval* (high and low bounding values), and *size* (number of data points in each bin). The output at the data point (areal unit) level is its assigned bin and the number of times it has fallen into this bin (and also other bins). Next, we place each areal unit in its most frequent *binID*, and subsequently detect and resolve conflicts (Lines 20 - 24).

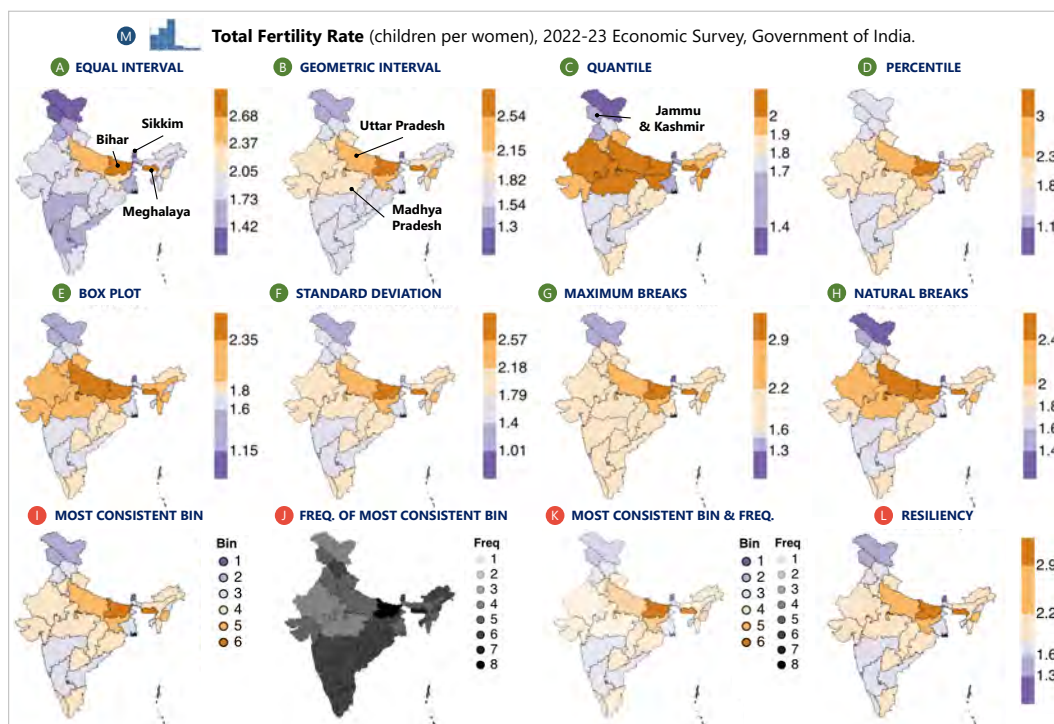
³ Binning methods are considered comparable if they have the same resultant (or specified) bin count, e.g., if the specified bin count is five, then we can compare *equal interval*, *quantile*, *maximum breaks*, *natural breaks*, *ck-means*, and *geometric interval*. If the desired bin count is six, then *box-plot* and *percentile* may also be considered (as they generally output six bins). Other methods may also be considered on a case-by-case basis, e.g., *defined interval* if the specified bin interval results in the desired bin count.

We note three possible conflicts. First, if there is a tie for the most frequent bin assignment, we break the tie on a first-come-first-serve basis and use the smaller *binID*. Next, resultant bin extents could overlap, e.g., *binID* = 1’s maximum extent is 50 and *binID* = 2’s minimum extent is 45 (smaller instead of greater); in response, we skip the most frequent *binID* assignment and iterate on the next (i.e., second) most frequently assigned bin and so on. Third, some bins (e.g., the middle bin) may have no areal units; one could output fewer bins or equally split bins until the desired bin count is reached; *resiliency* supports both modes.

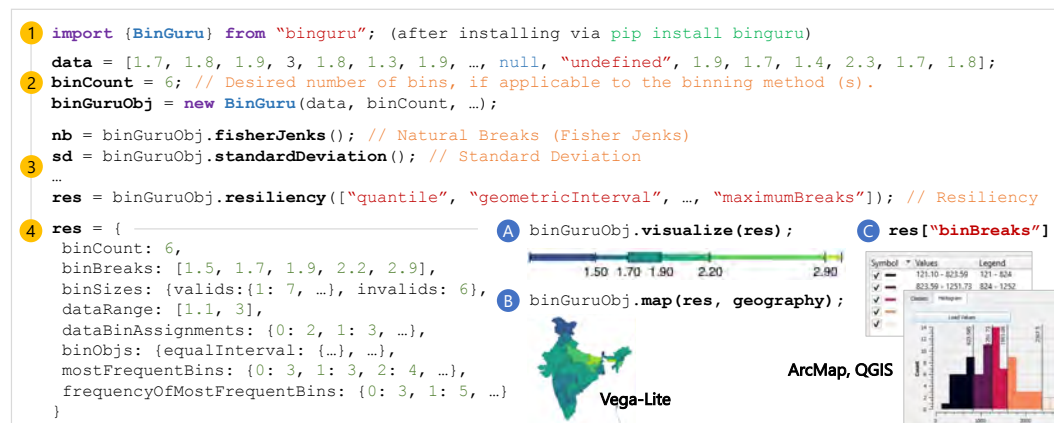
3.1 Usage Scenario

Imagine Kiran is an Indian government official who wants to map the “Total Fertility Rate” (number of children per women) across India to educate the general public and inform future birth-related policies. The data contains 28 states and 7 union territories, and the total fertility rate ranges from 1.1 (Sikkim) to 3.0 (Bihar) with an average of 1.8 children per woman. They are not sure which binning method to use.

Kiran uploads their dataset to an interactive notebook we developed powered by the *BinGuru* JavaScript library. They choose *six* bins and a divergent *purple to orange* color scheme, and inspect the output of eight (out of 18 supported) binning methods: *equal interval*, *geometric interval*, *quantile*, *maximum breaks*, *percentile*, *box plot*, *standard deviation*, and *natural breaks* (Figure 1A-H). They find that states of Bihar and Meghalaya (north east) have the largest fertility rates with *equal interval*, *geometric interval*, *maximum breaks*, and *standard deviation*. However, when using the *quantile*, *box plot*, and *natural breaks* methods, Uttar Pradesh (west of Bihar) is also placed in the same bin. *Maximum breaks* groups most



■ **Figure 1** Small multiples of choropleth maps showing “Total Fertility Rate (children per women)” (M) in India [8] using established binning methods (A-H) and *resiliency* (I-L).



■ **Figure 2** Usage scenario demonstrating how developers can (1) import the “binguru” library, (2) initialize a **BinGuru** class instance with the input data and binning parameters (e.g., **binCount** – desired number of bins), (3) explore different binning methods (e.g., **.fisherJenks()**), and (4) inspect their output comprising resultant *binBreaks* (bin boundaries), *binSizes* (number of points in each bin), *dataBinAssignments* (binID corresponding to each point), *binObjs* (applicable for *resiliency*, with intermediate binning outputs of constituent binning methods), *mostFrequentBins*, and *frequencyOfMostFrequentBins*. Developers can also visualize the output on (A) a legend, (B) a map – if underlying geography is available, and/or (C) copy-paste into commercial GIS software.

of western, central, and southern regions in the same bin (*binID* = 4). *Quantile* shows a more even distribution but does not distinguish extreme values. Kiran is uncertain which method to use and experiments with *resiliency*, visualizing the results using three maps (Figure 1I-L):

Most Consistent Bin. This map shows the most *frequent* bin across binning methods for each areal unit (Figure 1I). For example, Bihar (*Total Fertility Rate* equals 3.0) is colored dark orange, implying it is most frequently placed in *binID* = 6—with a high fertility rate.

Frequency of Most Consistent Bin. This map shows the *frequency of an areal unit’s most frequent (or consistent) bin* assignment across binning methods (Figure 1J). Higher numbers mean that the areal unit consistently fell into the same bin. For example, Bihar is colored the darkest shade of black (*frequency* = 8), implying it is consistently placed in the same bin (*binID* = 6); whereas, Madhya Pradesh (center) is colored much lighter, implying it is inconsistent across bins. Kiran understands that their binning decisions will affect how Madhya Pradesh is classified and will discuss this “fuzziness” at future meetings.

Most Consistent Bin and its Frequency. This bivariate map combines the previous two maps into a value-by-alpha map [13]. The hue corresponds to the *most frequent bin* and the opacity corresponds to the *frequency of the most frequent bin* (Figure 1K), where higher opacity implies higher frequency. Here, Bihar is an opaque orange color, as it consistently falls in *binID* = 6. More transparency indicates low frequency, less certainty, and inconsistency.

Resiliency. This map (Figure 1L) often (but not always) resembles Figure 1I, but now includes a legend that reflects the actual data values as the bin breaks. Kiran observes that *Resiliency* retained the regions of Bihar and Meghalaya as the regions with the largest and Sikkim with the smallest (outlier) values, while also showing variance among other northern, southern, and western states. They also note that this result resembles the *standard deviation*

method (Figure 1F) except for the bin assignment of the regions of Jammu & Kashmir (north) and Sikkim (north east). They now decide to either use the output of *resiliency* as-is or use *standard deviation*, which they value as a more familiar, easy-to-understand method.

4 Implementation, Future Work and Conclusion

Resiliency and 17 other binning methods are available through an open-source JavaScript library, **BinGuru** (Figure 2). We next plan to make *resiliency* more robust with weighting (e.g., *equal interval* has more weight). We then plan to better guide users by recognizing the distribution of their data and suggesting binning methods that are appropriate for that distribution. We also hope to capture how cartographers and GIS experts might use *resiliency* to learn about its benefits and drawbacks, ease of use, and uptake to drive future iterations.

In conclusion, we presented *resiliency*, a new data binning method that assigns areal units to their most agreed-upon, consensus bin as it persists across multiple binning methods. We showed how *resiliency* can enable spatial patterns to be communicated with(out) outliers and promote reflection on the fuzziness often imposed during binning.

References

- 1 Marc Armstrong, Ningchuan Xiao, and David Bennett. Using Genetic Algorithms to Create Multicriteria Class Intervals for Choropleth Maps. *Annals of the Association of American Geographers*, 93:595–623, September 2003.
- 2 Arnold Bregt and Marco CS Wopereis. Comparison of complexity measures for choropleth maps. *The Cartographic Journal*, 27(2):85–91, 1990.
- 3 Cynthia A. Brewer and Linda Pickle. Evaluation of Methods for Classifying Epidemiological Data on Choropleth Maps in Series. *Annals of the Association of American Geographers*, 92(4):662–681, 2002.
- 4 Michael John De Smith, Michael F Goodchild, and Paul Longley. *Geospatial Analysis: A Comprehensive Guide To Principles, Techniques And Software Tools*. Troubador Publishing Ltd., 2007.
- 5 ESRI. ArcGIS, 2023. URL: <https://www.esri.com/en-us/arcgis/about-arcgis/overview>.
- 6 ESRI. ArcGIS Maps SDK, 2023. URL: <https://developers.arcgis.com/javascript/latest/api-reference/>.
- 7 Alan M MacEachren. Map complexity: Comparison and measurement. *The American Cartographer*, 9(1):31–46, 1982.
- 8 Ministry of Finance, Government of India. Economic Survey of India, 2023. URL: <https://www.indiabudget.gov.in/economicsurvey/doc/Statistical-Appendix-in-English.pdf>.
- 9 Mark Monmonier. Maximum-Difference Barriers: An Alternative Numerical Regionalization Method. *Geographical Analysis*, 5(3):245–261, 1973.
- 10 Mark Monmonier. Flat laxity, optimization, and rounding in the selection of class intervals. *Cartographica: The International Journal for Geographic Information and Geovisualization*, 19(1):16–27, 1982.
- 11 Mark Monmonier. *How To Lie With Maps*. University of Chicago Press, 2018.
- 12 Serge Rey and Luc Anselin. PySAL, 2005. URL: <https://pysal.org/>.
- 13 Robert E Roth, Andrew W Woodruff, and Zachary F Johnson. Value-By-Alpha Maps: An Alternative Technique To The Cartogram. *The Cartographic Journal*, 47(2):130–140, 2010.
- 14 Vidya Setlur, Michael Correll, and Sarah Battersby. Oscar: A semantic-based data binning approach. In *2022 IEEE Visualization and Visual Analytics (VIS)*, pages 100–104, Los Alamitos, CA, USA, October 2022. IEEE Computer Society.
- 15 Gary Sherman. QGIS, 2002. URL: <https://qgis.org/>.

- 16 Terry A Slocum, Robert B McMaster, Fritz C Kessler, and Hugh H Howard. *Thematic Cartography And Geovisualization*. CRC Press, 2022.
- 17 Richard M Smith. Comparing traditional methods for selecting class intervals on choropleth maps. *The Professional Geographer*, 38(1):62–67, 1986.
- 18 tmap. Tmap, 2023. URL: <https://cran.r-project.org/web/packages/tmap>.
- 19 Waldo R Tobler. Choropleth Maps Without Class Intervals? *Geographical Analysis*, 1973.

Counter-Intuitive Effect of Null Hypothesis on Moran's I Tests Under Heterogenous Populations

Hayato Nishi¹   

Graduate School of Social Data Science, Hitotsubashi University, Tokyo, Japan

Ikuho Yamada   

Center for Spatial Information Science, The University of Tokyo, Japan

Abstract

We examine the effect of null hypothesis on spatial autocorrelation tests using Moran's I statistic. There are two possible variable states that do not exhibit spatial autocorrelation. One is that they have the same average values in all small regions, and the other is that they are not the same, but their variations are spatially random. The second state is less restrictive than the first. Thus, it intuitively appears suitable for the null hypothesis of Moran's I test. However, we found that it can make false discoveries more frequently than the nominal rate of the test when the first state is the true data generation process.

2012 ACM Subject Classification Information systems → Geographic information systems

Keywords and phrases Moran's I statistic, spatial autocorrelation, spatial heterogeneity, false discovery, null hypothesis

Digital Object Identifier 10.4230/LIPIcs.GIScience.2023.56

Category Short Paper

Funding *Hayato Nishi*: CSIS Joint Research Grants Program, Center for Spatial Information Science, the University of Tokyo.

Ikuho Yamada: JSPS KAKENHI Grant Number JP 22H00245.

1 Introduction

Moran's I statistic [3] is one of the most widely accepted statistics for testing spatial autocorrelation in spatially aggregated quantitative data such as the results of social surveys aggregated at the municipality level. A typical example of data to be tested is “per capita” quantity. For instance, we may obtain the average income of each municipality from a survey and test whether spatial clusters of high (or low) income exist using these data. In this paper, we discuss two fundamental aspects of Moran's I test that are often overlooked but can potentially affect the results of the test. One is the reliability of the observations and the other is the null hypothesis.

The reliability of the observations varies among municipalities because of their heterogeneous populations and sizes. Although the original implementation of Moran's I test does not consider such variability in data reliability, studies have pointed out its influence on results and proposed adjustment methods for heterogeneous populations [4, 7, 1].

In addition to population heterogeneity, the selection of the null hypothesis also affects the results of Moran's I test. [1] classified the spatial risk pattern (which corresponds to the income pattern in our example) to be tested into three states:

- A . spatially constant risk,
- B . heterogeneous risks without spatial correlation, and
- C . heterogeneous risks with spatial correlation.

¹ Corresponding author



Although the Hypotheses \mathcal{A} and \mathcal{B} imply no spatial autocorrelation, their practical meanings are substantially different. Hypothesis \mathcal{A} is rejected when there are differences in the average income of individual municipalities. By contrast, \mathcal{B} is rejected only when the differences in average income have spatial clusters. Therefore, we consider \mathcal{A} as a more rigorous state of no spatial autocorrelation than \mathcal{B} . When one suspects that the data in hand have a spatial pattern of \mathcal{C} , it appears reasonable to employ \mathcal{B} as the null hypothesis to detect spatial autocorrelation in the data. Employing \mathcal{A} as the null hypothesis would result in over-detection because it regards the spatial pattern of \mathcal{B} as spatial autocorrelation. However, analysts do not always carefully examine the null hypothesis when applying Moran's I test. In this study, we investigate how our choice of null hypothesis and population adjustment influences the results of Moran's I test.

This paper is structured into four sections, including this introduction. Section 2 discusses the theoretical basis for adjusting Moran's I test for heterogeneous populations. Section 3 presents simulation studies using synthetic grids and population data for Japanese municipalities. Section 4 summarizes our major findings.

2 Spatial Autocorrelation Tests with Moran's I

2.1 Moran's I Statistic

Let us consider a set of observed values $\mathbf{x} = (x_1, \dots, x_n)^\top$ for a study region consisting of n regions. Let \mathbf{C} be a known spatial adjacency matrix and $c_{i,j}$ be its $i-j$ element. When regions i and j are adjacent, $c_{i,j} = 1$; otherwise, $c_{i,j} = 0$. Furthermore, for diagonal elements, $c_{i,i} = 0$. Let \mathbf{W} be a row-standardized version of \mathbf{C} and $w_{i,j}$ be the $i-j$ element. In the simulation studies discussed in Section 3, we define \mathbf{C} as Queen's contiguity matrix. Using these notations, Moran's I statistic is defined as

$$I(\mathbf{x}) = \frac{n\mathbf{x}^\top \mathbf{M} \mathbf{W} \mathbf{M} \mathbf{x}}{W_0 \mathbf{x}^\top \mathbf{M} \mathbf{x}} \quad (1)$$

where $W_0 = \sum_i \sum_j w_{i,j}$ and $\mathbf{M} = \mathbf{I} - \frac{1}{n} \mathbf{1} \mathbf{1}^\top$. Note that \mathbf{I} is the identity matrix of size n and $\mathbf{1}$ is an $n \times 1$ vector, all of whose elements are 1.

2.2 Data Generation Process and Null Hypothesis

Here, we derive the distribution of Moran's I when \mathbf{x} follows the Gaussian distribution. We assume that x_i represents the estimated value of an unknown parameter μ_i . For instance, let x_i be the average income observed in region i , μ_i be its true value without biases such as measurement errors, and $y_{i,k}$ be income that an individual k in region i gains. As $y_{i,k}$ generally contains personal differences and measurement errors, we assume that $y_{i,k}$ follows a normal distribution with mean μ_i and variance σ^2 . Letting m_i be the population of region i , x_i is given by $\frac{1}{m_i} \sum_k y_{i,k}$; thus, it can be discerned that the observation x_i follows a normal distribution with mean μ_i and variance $\frac{\sigma^2}{m_i}$. If the data generation process (DGP) is \mathcal{A} , the mean μ_i is constant μ for the entire study region. However, if DGP is \mathcal{B} , μ_i is not uniform. Following [1], we assume that μ_i follows an independent normal distribution of mean μ and variance $\sigma^2 s^2$. The parameter s^2 controls the relative heterogeneity of true values μ_i . If $s^2 = 0$, then the DGP corresponds to \mathcal{A} , whereas if $s^2 > 0$, it corresponds to \mathcal{B} .

Therefore, letting $\boldsymbol{\mu} = (\mu_1, \dots, \mu_n)^\top$ be the vector of true income values and $\boldsymbol{\Sigma}$ be a diagonal matrix whose $i-i$ element is $\frac{1}{m_i} + s^2$, \mathbf{x} follows a multivariate normal distribution of the mean $\boldsymbol{\mu}$ and the variance-covariance matrix $\boldsymbol{\Sigma}$. Below we explain our finding that,

when the mean μ_i is constant μ for the entire study region, and $s^2 = 0$, the distribution of Moran's I does not depend on unknown parameters μ and σ^2 . When Σ can be decomposed into $\Sigma = \mathbf{L}\mathbf{L}^\top$ by Cholesky decomposition,

$$\mathbf{x} = \mu\mathbf{1} + \sigma\mathbf{L}\boldsymbol{\varepsilon} \tag{2}$$

where $\boldsymbol{\varepsilon}$ is a vector of elements following a standard normal distribution. By substituting this into x in Eq. (1), we can obtain

$$I(\mathbf{x}) = \frac{n\mathbf{x}^\top\mathbf{M}\mathbf{W}\mathbf{M}\mathbf{x}}{W_0\mathbf{x}^\top\mathbf{M}\mathbf{x}} = \frac{n\boldsymbol{\varepsilon}^\top\mathbf{L}^\top\mathbf{M}\mathbf{W}\mathbf{M}\mathbf{L}\boldsymbol{\varepsilon}}{W_0\boldsymbol{\varepsilon}^\top\mathbf{L}^\top\mathbf{M}\mathbf{x}} \tag{3}$$

given that $\mathbf{M}\mathbf{L} = \mathbf{0}$ and $\mathbf{M}\mathbf{x} = \mu\mathbf{M}\mathbf{1} + \sigma\mathbf{M}\mathbf{L}\boldsymbol{\varepsilon} = \sigma\mathbf{M}\mathbf{L}\boldsymbol{\varepsilon}$, where $\mathbf{0}$ is a zero vector. Eq. (3) includes neither the parameters μ nor σ^2 , implying that Moran's I statistic is a pivotal statistic independent of the unknown parameters when we assume \mathcal{A} as a null hypothesis. [5] and [6] present the distribution of Moran's I statistic and its approximation, respectively, when the observed vector \mathbf{x} follows a normal distribution. Based on them and Eq. (3), the probability that $I(x)$ is less than an arbitrary value I_{obs} can be written as

$$\Pr [I(x) \leq I_{obs}] = \Pr [\boldsymbol{\varepsilon}^\top (n\mathbf{L}^\top\mathbf{M}\mathbf{W}\mathbf{M}\mathbf{L} - I_{obs}W_0\mathbf{L}^\top\mathbf{M}\mathbf{L}) \boldsymbol{\varepsilon} \leq 0]. \tag{4}$$

Let \mathbf{T} be $n\mathbf{L}^\top\mathbf{M}\mathbf{W}\mathbf{M}\mathbf{L} - I_{obs}W_0\mathbf{L}^\top\mathbf{M}\mathbf{L}$ and its eigenvalue decomposition be $\mathbf{T} = \mathbf{E}^\top\boldsymbol{\Lambda}\mathbf{E}$, where $\boldsymbol{\Lambda}$ is a diagonal matrix composed of the eigenvalues, $(\lambda_1, \dots, \lambda_n)$. If we make \mathbf{E} an orthogonal matrix, $\mathbf{z} = \mathbf{E}\boldsymbol{\varepsilon}$ follows independent normal distributions; thus, the left-hand side of the inequality in Eq. (4), $\boldsymbol{\varepsilon}^\top\mathbf{T}\boldsymbol{\varepsilon} = \sum_i \lambda_i z_i^2$, follows the generalized chi-square distributions. [2] provides details of this transformation. This property indicates that we can evaluate the cumulative distribution of Moran's I statistic by evaluating that of the generalized chi-square distribution without using the unknown parameters μ and σ^2 .

This property is particularly beneficial when the population m_i is not uniform because we cannot employ the permutation test approach because the observation vector \mathbf{x} is not exchangeable. If our null hypothesis is \mathcal{A} , then we assume $s^2 = 0$. Hence, we can apply population adjustment without knowledge of the unknown parameters μ and σ^2 . However, if our null hypothesis is \mathcal{B} , we need s^2 for the population adjustment.

We cannot distinguish \mathcal{A} and \mathcal{B} when the population m_i is uniform for the entire study region. Therefore, the selection of the null hypothesis \mathcal{A} or \mathcal{B} does not affect the property of Moran's I test when the populations are uniform and the DGP is Gaussian. By contrast, in the case of heterogeneous populations, it is unclear how the selection of the null hypothesis affects the results. In the next section, we examine the potential influence of this selection using two simulation studies.

3 Simulation Studies

This section describes the settings and results of the simulations. The two study regions are discussed in Sections 3.1 and 3.2. Section 3.1 presents a synthetic grid system with three population patterns to examine the influence of population heterogeneity. For a more realistic scenario, we introduce real-world municipalities and their populations in Section 3.2. Section 3.3 illustrates how the choice of null hypothesis influences the false discovery rate (FDR) of Moran's I test in our simulations.

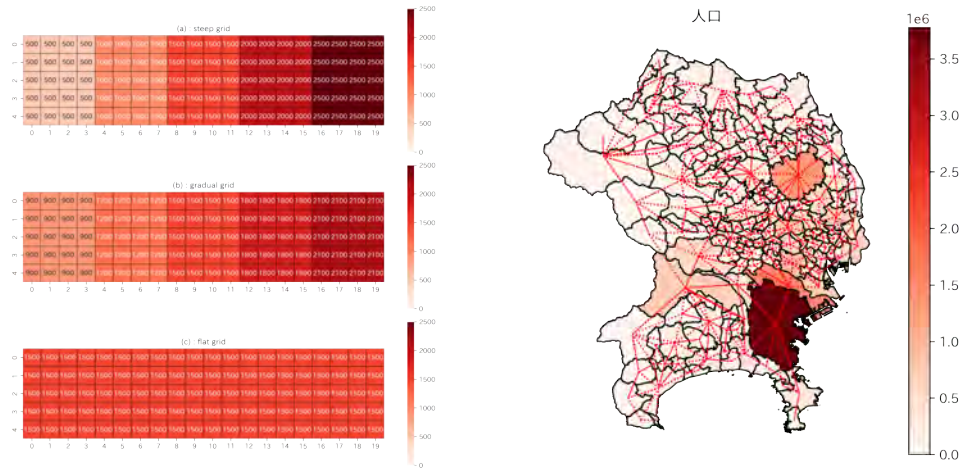


Figure 1 Populations on the Synthetic Grids.

Figure 2 Municipalities and Populations in Tokyo, Japan.

3.1 Synthetic Grid Data

We consider a 20×5 regular grid system as the study region. Using the notation defined in Section 2 and assuming that an individual living in region i has the value of a target variable with mean μ_i and variance σ^2 , the value of x_i can be simulated as a random number obtained from the normal distribution of mean μ_i and variance $\frac{\sigma^2}{m_i}$. Note that variance σ^2 is set constant for the entire study region. Once the local mean μ_i is marginalized, the observation x_i follows a normal distribution of the mean μ and variance $\sigma^2(\frac{1}{m_i} + s^2)$.

To examine the influence of heterogeneous populations, we consider the three spatial distributions of the regional populations shown in Figure 1. The “steep grid” pattern shown in Figure 1(a) has the regional population that steeply increases toward the right, while the “flat grid” pattern in Figure 1(c) shows a constant regional population for the entire study region. The “gradual grid” pattern in Figure 1(b) is in between; while its regional population also increases toward the right, it is less steep than the steep grid pattern. The regional populations are arranged such that the total populations are the same.

In the simulations described in Section 3.3, we set $\sigma^2 = 1.0$ and $\mu = 0$. For the nonspatial autocorrelation state \mathcal{B} , we select $s^2 = 1.0$.

3.2 Tokyo Municipality Data

For a realistic study region, we use municipal and population data from three prefectures in the Tokyo Metropolitan Area in Japan: Tokyo, Kanagawa, and Saitama. The municipal boundaries and populations are shown in Figure 2. The red dashed lines show the neighborhood relationships among municipalities defined by the Queen style.

In the simulations in Section 3.3, we set σ^2 as the same as the average of m_i and $\mu = 0$. For the nonspatial autocorrelation state \mathcal{B} , we select $s^2 = \sigma^{-2}$.

3.3 Results

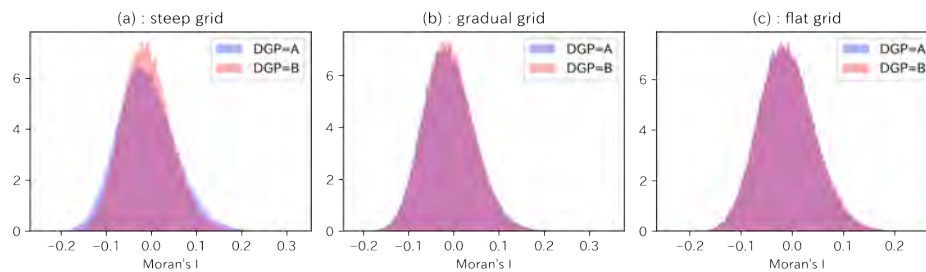
For both the grids and Tokyo, we applied one-sided tests to detect a positive autocorrelation at the 5% significance level. We employed the numerical approach presented in [5] to calculate the cumulative probability that appears in Eq. (4). Therefore, numerical errors were included in the simulation results.

■ **Table 1** False Discovery Rates on the Synthetic Grids.

(a) steep grid		
	$H_0 = \mathcal{A}$	$H_0 = \mathcal{B}$
DGP= \mathcal{A}	0.049	0.077
DGP= \mathcal{B}	0.027	0.049
(b) gradual grid		
	$H_0 = \mathcal{A}$	$H_0 = \mathcal{B}$
DGP= \mathcal{A}	0.050	0.057
DGP= \mathcal{B}	0.043	0.049
(c) flat grid		
	$H_0 = \mathcal{A}$	$H_0 = \mathcal{B}$
DGP= \mathcal{A}	0.050	0.050
DGP= \mathcal{B}	0.050	0.050

■ **Table 2** False Discovery Rates on Tokyo Municipalities.

	$H_0 = \mathcal{A}$	$H_0 = \mathcal{B}$
DGP= \mathcal{A}	0.050	0.058
DGP= \mathcal{B}	0.042	0.049



■ **Figure 3** The Distributions of Moran's I on the Synthetic Grids.

Table 1 shows the false discovery rates (FDR) of the synthetic grids described in Section 3.1. In the case of (c) flat grid, \mathcal{A} and \mathcal{B} are identical, as discussed in Section 2.2. Thus, we do not need to consider differences in the null hypothesis if the population is uniform. However, in other grids, FDRs equal to a nominal rate of 5% only when the null hypothesis H_0 is correctly selected. This shows that the null hypothesis $H_0 = \mathcal{B}$, which allows heterogeneity of the true mean μ_i , results in a much higher FDR than expected, when actual μ_i is homogeneous. The opposite result is obtained when we employ $H_0 = \mathcal{A}$. Thus, counterintuitively, a test that assumes homogeneous means is more conservative than one that allows heterogeneous means. This tendency is clearer in (a) steep grid than in (b) gradual grid.

Table 2 shows the result of Tokyo municipality data. We observe the same counterintuitive results as those found in synthetic grids.

The results in Tables 1 and 2 indicate that $H_0 = \mathcal{A}$ is a safer choice than $H_0 = \mathcal{B}$ to keep FDR less than 5%, which is the predetermined nominal significance level of the test. This is because the distribution of Moran's I from $H_0 = \mathcal{A}$ exhibits a larger variance than from $H_0 = \mathcal{B}$. Figure 3 shows Moran's I distributions for the synthetic grids. However, whether this property is always observed remains unclear.

4 Conclusion

Intuitively, the test under the null hypothesis \mathcal{B} does not reject it if the true data generation process (DGP) is \mathcal{A} . Hence, it sounds reasonable for analysts to employ \mathcal{B} as their null hypothesis if they want to discover only \mathcal{C} . However, our simulation studies based on

synthetic grids and real municipalities with population data revealed that testing under the null hypothesis \mathcal{B} does not guarantee that FDR becomes less than the nominal significance level if the true DGP is \mathcal{A} . In other words, if we employ \mathcal{B} as a null hypothesis, we may often detect incorrect “spatial autocorrelation” of income when income is the same in all municipalities. This implies that the null hypothesis must be selected carefully when applying spatial autocorrelation test.

Further research is needed to examine whether this counterintuitive property appears in other situations, such as the target variable x_i following non-Gaussian distributions and the spatial contiguity matrix \mathbf{C} different from Queen's definition. To evaluate the performance of the test, the statistical power, in addition to FDR, also needs to be examined. This is not straightforward because the true value of s^2 is generally unknown; thus, practical approaches are required.

References

- 1 Renato M. Assunção and Edna A Reis. A new proposal to adjust Moran's I for population density. *Statistics in Medicine*, 18(16):2147–2162, 1999.
- 2 Abhranil Das and Wilson S. Geisler. A method to integrate and classify normal distributions. *Journal of Vision*, 21(10):1, September 2021.
- 3 P. A. P. Moran. Notes on Continuous Stochastic Phenomena. *Biometrika*, 37(1/2):17, June 1950.
- 4 Neal Oden. Adjusting Moran's I for population density. *Statistics in Medicine*, 14(1):17–26, January 1995.
- 5 Michael Tiefelsdorf. Some practical applications of Moran's I's exact conditional distribution. *Papers in Regional Science*, 77(2):101–129, 1998.
- 6 Michael Tiefelsdorf. The saddlepoint approximation of Moran's I's and local Moran's Ii's reference distributions and their numerical evaluation. *Geographical Analysis*, 34(3):187–206, 2002.
- 7 Thomas Waldhör. The spatial autocorrelation coefficient Moran's I under heteroscedasticity. In *Statistics in Medicine*, volume 15, pages 887–892, 1996.

A Data Fusion Framework for Exploring Mobility Around Disruptive Events

Evgeny Noi¹  

Department of Geography, University of California Santa Barbara, CA, USA

Somayeh Dodge 

Department of Geography, University of California Santa Barbara, CA, USA

Abstract

This paper proposes a data fusion framework that seeks to investigate joint mobility signals around wildfires in relation to geographic scale of analysis (level of spatial aggregation), as well as spatial and temporal extents (i.e. distance to the event and duration of the observation period). We highlight the usefulness of our framework using intra-urban mobility data from Mapbox and SafeGraph for two wildfires in California: Lake Fire (August-September 2020, Los Angeles County) and Silverado Fire (October-November 2020, Orange County). We identify two distinct patterns of mobility behavior: one associated with the wildfire event and another one - with the routine daily mobility of the nearby urban core. Using the combination of data fusion and tensor decomposition, the framework allows us to capture additional insights from the data, that were otherwise unavailable in raw mobility data.

2012 ACM Subject Classification Information systems → Geographic information systems

Keywords and phrases geographic extent, geographic scale, tensor decomposition, spatio-temporal analysis

Digital Object Identifier 10.4230/LIPIcs.GIScience.2023.57

Category Short Paper

Funding *Somayeh Dodge*: NSF Award # 2043202: Modeling Movement and Behavior Responses to Environmental Disruptions.

Acknowledgements The authors gratefully acknowledge the support from the National Science Foundation through award BCS # 2043202. Mobility data provided by ©MapBox and ©SafeGraph.

1 Introduction

The issue of geographic scale (i.e. level of detail / aggregation), geographic extent (i.e. area of analysis, measured in terms of proximity to the phenomena of interest), and temporal extent (i.e. the period of observation in relation to the phenomena of interest) has been an important research topic in GIScience and in movement research within the last decade [5, 3]. Many conventional methods of geographic analysis are traditionally designed for univariate spatial or temporal series (e.g. geographically-weighted regression, local indicators of spatial association, spatial scan statistics). Yet, data on human movement is increasingly heterogeneous [1, 6], multivariate, and dependent on local land-use and transportation patterns, necessitating further development of complex multivariate spatio-temporal methods that can leverage and integrate numerous data sources.

We propose an analytical framework that allows to fuse various indicators of human mobility (in this paper, only two are considered) at different geographic and temporal scales to identify the impact zone of disruptive events, such as wildfires, on mobility. The framework consists of several processes: 1) Multi-scale spatio-temporal matching of data to

¹ corresponding author



combine various types of geographic data (e.g. POI point vector data and regular grid-based OpenStreetMap tiles) geographically and temporally. 2) Calculating mutual information score combining the multiple data sets for different geographic and temporal scale and extent. 3) Fitting the fused data to PARAFAC2 tensor decomposition model [4] to elicit shared patterns of movement at different locations. As a case study, we test this framework for exploring mobility patterns around two wildfire events.

2 Methods

2.1 Multi-Scale Spatio-Temporal Matching and Aggregation

The first step in the proposed framework relies on the matching of the various data sources (in this case two). These data sources vary in coverage, have different units of analysis and units of measurement. The process of matching is described in Algorithm 1. In short, it sequentially increases the spatial scale of the study area (unit of analysis), the spatial extent (e.g. distance from the fire event), and the temporal extent of the observations (e.g. time from fire ignition) to fuse the two mobility indices into a mutual information (MI) score as described in the next section (see Figure 1c).

Algorithm 1 Multi-Scale Spatio-Temporal Matching.

Input : fire perimeter, mobility index 1 (M_1), mobility index 2 (M_2), spatial extent (distance from fire perimeter - S_i), temporal extent (days from the fire ignition - T_j), spatial scale / zoom levels (OpenStreetMap zoom level tiles - Z_k)

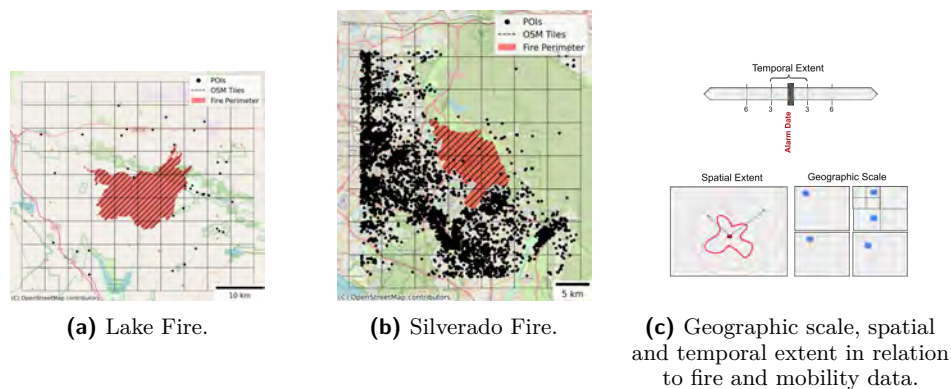
Output : Mutual information MI_{zst} for various spatial extent ($S \in 1\dots j$), temporal extent ($T \in 1\dots j$) and spatial scale ($Z \in 1\dots k$)

- 1 Discretize study area using OSM tiles at level Z_k and aggregate mobility indices (M_1, M_2) at selected spatial S_i and temporal T_j extent levels;
- 2 for $z \leftarrow 1$ to k do
- 3 for $t \leftarrow 1$ to j do
- 4 for $s \leftarrow 1$ to i do
- 5 1. Keep only spatial units that have non null values across M_1 and M_2 ;
- 6 2. Calculate mutual information MI_{zst} from M_{1zst} and M_{2zst} ;
- 7 end
- 8 end
- 9 end

2.2 Mutual Information (MI)

Mutual information (MI) is a measure of the amount of information that two variables share. In information theory, MI is defined as the reduction in uncertainty about one variable (X) given knowledge of another variable (Y). In contrast to Pearson correlation, the MI score is ideally suited to capture non-linear dependence between random variables. Mathematically, the MI between two discrete random variables X and Y is defined as: $MI(X; Y) = H(X) - H(X|Y)$, where $H(X)$ is the entropy of X , which measures the amount of uncertainty in X , and $H(X|Y)$ is the conditional entropy of X given Y , which measures the remaining uncertainty in X when Y is known.

The rationale behind utilizing the mutual information of mobility is simple: the premise is that in the presence of emergency events such as a natural disaster, all types of mobility are affected. As such, variation in mobility will be manifested in various measured indices of mobility. By utilizing mutual information across different geographical scales and spatio-temporal extents, we hope to establish and characterize mutual dependence of mobility indices and fuse movement data that may be different in coverage, uncertainty, and bias.



■ **Figure 1** POI location and OSM grid representation of the study areas. For demonstration purpose only the zoom level 13 is illustrated, denoting the coarsest level of detail.

2.3 Tensor Decomposition

Tensors are multi-dimensional arrays and, as such, require multi-dimensional methods of analysis. Tensor decomposition allows us to uncover hidden latent factors (clusters of behavior) in multi-dimensional data. There are different types of tensor decomposition, including Tucker, CANDECOMP, and Tensor-Train [7]. Of particular interest to this study is the PARAFAC2 factorization [4], because it allows to jointly model data arrays of different sizes (for instance, where the spatial extent of data varies), by aligning them across a shared dimension (e.g. time). The multiset data can be decomposed as follows: $\mathbf{X}_k = \mathbf{U}_k \mathbf{S}_k \mathbf{V}^T$, where R is the number of components derived from the decomposition, $\mathbf{U}_k \in \mathbb{R}^{I_k \times R}$, $\mathbf{S}_k \in \mathbb{R}^{R \times R}$, and $\mathbf{V} \in \mathbb{R}^{J \times R}$. The PARAFAC2 decomposition is fitted via an alternating direction method of multipliers (AO-ADMM) [9] available through MatCouply Python package [8]. Since MI scores are always non-negative, we impose non-negativity constraints on the uncovered decomposition components (factors).

3 Case Study

3.1 Study Area

This study focuses on two major wildfires in California (Figure 1a, 1b) that happened in 2020: Lake Fire, which burned around 31,000 acres in Angeles National Forest in Los Angeles County from August 12 to September 28 and Silverado Fire, which burned around 13,000 acres in Orange County from October 26 to November 7, 2020. The data for this study was collected specifically for two months before and after the ignition date of the two wildfires: Silverado Fire (August 26 - December 26, 2020) and Lake Fire (June 12 - October 12, 2020).

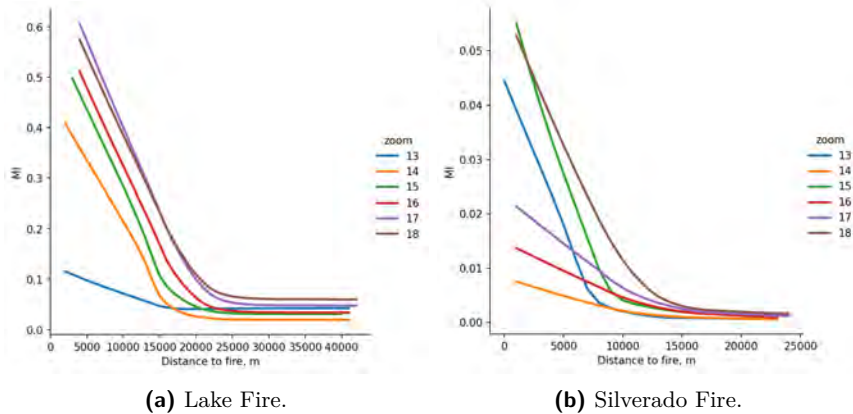
3.2 Data

SafeGraph published several mobility data products during the COVID-19 pandemic through their Data for Good Initiative. One of such products is *Weekly Patterns*, which reports raw visitor counts at the points of interest (POI) level daily. Mapbox provides gridded data, representing the amount of mobility, at OpenStreetMap (OSM) tile level (with OSM zoom level 18, finest resolution available, corresponding to a square grid of $100\text{m} \times 100\text{m}$). The data is aggregated and delivered daily, and is available in the form of an activity index, ranging from 0 to ∞ , where higher index values denote higher levels of mobility.

To investigate the mutual information at various levels of aggregation using the proposed framework, we utilize OSM zoom levels 13–18, where zoom level 13 corresponds to a 1:70,000 screen scale (village level), and level 18 corresponds to a 1:2,000 scale (buildings/trees levels)². OpenStreetMap tiles are regular square tessellations that remain uniform across remote and isolated areas where wildfires occur. Thus we can ascertain mobility at different spatial scales, while minimizing modifiable areal unit problem [2]. To delineate the study area, we create a bounding box from a 10km buffer around each of the wildfire perimeters and filter the mobility data to this extent. We hypothesize that direct impact zone of wildfires will be pronounced the most in close vicinity to the fire.

3.3 Data Tensorization

The mutual information values are shaped into a multiset data \mathbf{X}_F , where $\mathbf{X} \in \mathbb{R}^{I \times J}$ is a matrix with spatial extent on the rows (I), temporal extent on the columns (J), and F denotes a fire name (in Algorithm 1). The bins for spatial extent are calculated starting from the centroid of the fire perimeter, and are incrementally increased by 1km, resulting in a progressively expanding geographical area around the fire. The distance of 1km provides a balanced binning for remote areas, when mobility data is sparse. The temporal extent (T_j) is measured in the number of days before and after the fire. For instance, if $j = 3$ we have a period of 6 days, starting 3 days before fire and terminating 3 days after the fire ignition. These bins are progressively increased by the increment of 3 to the total of 21 bins (i.e. 62 days or roughly two months). Thus, the final dimensions of the multiset data are as follows: $\mathbf{X}_{\text{lake}} \in \mathbb{R}^{43 \times 21}$ and $\mathbf{X}_{\text{silverado}} \in \mathbb{R}^{25 \times 21}$. Since the fire perimeters vary in size and shape, buffering and discretizing the study area will result in different number of spatial bins (\mathbf{X} rows): 43 rows for the Lake fire and 25 rows for the Silverado fire.



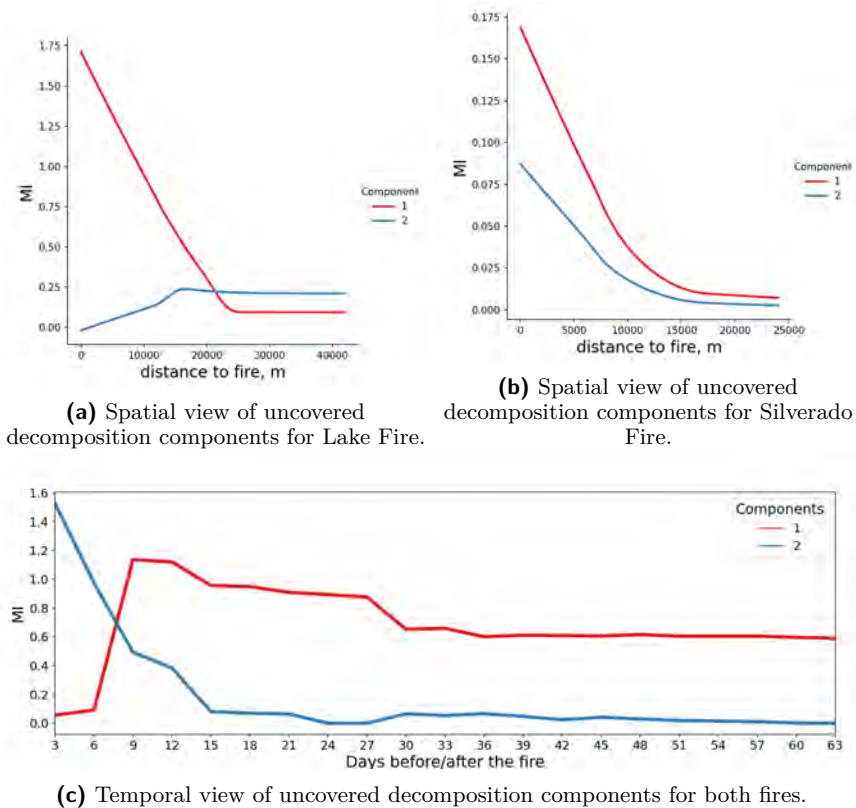
■ **Figure 2** Mutual information score curves for different levels of analysis (zoom levels) and geographical extent (radii from the fire).

4 Results

The fitted curves of the matched mutual information scores are plotted against the distance to the fire in Figure 2. For both fires the mutual information score decreases across various zoom levels as the distance from the fire increases. This is logical: as we include more spatial

² For more details see https://wiki.openstreetmap.org/wiki/Zoom_levels

units into our area of interest, we are also capturing daily urban mobility signal and noise. Since Lake Fire perimeter acreage is higher, the curves plateau around 20km from the fire (Figure 2a). One important difference between the two sets of curves is the magnitude of the fitted curves (noted in the different y-axis) which is largely due to drastically different number of POI at the two locations: so much so that the the MI scores are dominated by daily mobility, and not wildfire related mobility. This is supported by lack of relationship between temporal extent and MI scores.



■ **Figure 3** PARAFAC2 modeling results.

PARAFAC2 decomposition allows us to analyze both fires jointly, identifying distinct movement patterns (Figure 3) that are shared across two wildfires. A model with two decomposition components ($R = 2$) fits the data very well (99% fit). *Component 1* (denoted in red) shows fire-related mobility signal and plots MI scores against the spatial extent (Figure 3a, 3b). As we increase the geographic extent, the mutual information decreases for both fires (with more rapid decrease for sparse Lake Fire data), pointing to the higher dependence of mobility indices in close proximity to the fire event. On the temporal view for *Component 1* (Figure 3c) we notice that the MI scores are highest for the observation period of 9-12 days before/after the ignition date of the fire, declining gradually. This might be an indication that we need a relatively extended period of time to establish the effect of the wildfire. *Component 2* (denoted in blue) shows daily mobility associated with nearby urban cores at two locations. For Lake Fire (Figure 3a) the closest urban area, Lancaster is located approximately 15km to the Northwest of the fire (peak for blue line). For Silverado Fire (Figure 3b) the urban area borders with the fire perimeter on the Southeast, and as such, coincides with direct impact zone of the wildfire. On the temporal view for *Component 2*, the

MI scores fall abruptly, approaching zero around 15 days before/after the fire. This is logical, as we increase the observation period for two urban areas, there is no added information about the fire event.

5 Conclusion

This paper demonstrated how the proposed framework can be used to fuse the data on mobility at different spatial and temporal scale and establish relationships between mutual dependence of mobility indices around disruptive events. The mutual information score coupled with tensor decomposition is able to identify two clusters of behavior, which were otherwise not traceable in raw mobility signals (e.g. SafeGraph visitor counts or Mapbox activity index). The framework presented in this paper can be easily scaled up to incorporate more locations, different event types (hurricanes, floods, etc.) and event duration, and mobility indices. Future work will compare the methods laid out in the paper to other clustering techniques for studying aggregate human movement.

References

- 1 Somayeh Dodge. A data science framework for movement. *Geographical Analysis*, 53(1):92–112, 2021. doi:10.1111/gean.12212.
- 2 A Stewart Fotheringham and David WS Wong. The modifiable areal unit problem in multivariate statistical analysis. *Environment and planning A*, 23(7):1025–1044, 1991. doi:10.1068/a231025.
- 3 Michael F Goodchild. A GIScience perspective on the uncertainty of context. *Annals of the American Association of Geographers*, 108(6):1476–1481, 2018. doi:10.1080/24694452.2017.1416281.
- 4 Henk AL Kiers, Jos MF Ten Berge, and Rasmus Bro. PARAFAC2—Part I. A direct fitting algorithm for the PARAFAC2 model. *Journal of Chemometrics: A Journal of the Chemometrics Society*, 13(3-4):275–294, 1999. doi:10.1002/(SICI)1099-128X(199905/08)13:3/4%3C275::AID-CEM543%3E3.0.CO;2-B.
- 5 Mei-Po Kwan and Tijs Neutens. Space-time research in GIScience. *International Journal of Geographical Information Science*, 28(5):851–854, 2014. doi:10.1080/13658816.2014.889300.
- 6 Evgeny Noi, Alexander Rudolph, and Somayeh Dodge. Assessing COVID-induced changes in spatiotemporal structure of mobility in the United States in 2020: a multi-source analytical framework. *International Journal of Geographical Information Science*, 36(3):585–616, 2022. doi:10.1080/13658816.2021.2005796.
- 7 Evangelos E Papalexakis, Christos Faloutsos, and Nicholas D Sidiropoulos. Tensors for data mining and data fusion: Models, applications, and scalable algorithms. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 8(2):1–44, 2016. doi:10.1145/2915921.
- 8 Marie Roald. MatCoupLy: Learning coupled matrix factorizations with Python. *SoftwareX*, 21:101292, 2023. doi:10.1016/j.softx.2022.101292.
- 9 Marie Roald, Carla Schenker, Vince D Calhoun, Tulay Adali, Rasmus Bro, Jeremy E Cohen, and Evrim Acar. An AO-ADMM approach to constraining PARAFAC2 on all modes. *SIAM Journal on Mathematics of Data Science*, 4(3):1191–1222, 2022. doi:10.1137/21M1450033.

Finding Feasible Routes with Reinforcement Learning Using Macro-Level Traffic Measurements

Mustafa Can Ozkan¹ ✉ 

SpaceTimeLab, University College London, UK

Tao Cheng ✉

SpaceTimeLab, University College London, UK

Abstract

The quest for identifying feasible routes holds immense significance in the realm of transportation, spanning a diverse range of applications, from logistics and emergency systems to taxis and public transport services. This research area offers multifaceted benefits, including optimising traffic management, maximising traffic flow, and reducing carbon emissions and fuel consumption. Extensive studies have been conducted to address this critical issue, with a primary focus on finding the shortest paths, while some of them incorporate various traffic conditions such as waiting times at traffic lights and traffic speeds on road segments. In this study, we direct our attention towards historical data sets that encapsulate individuals' route preferences, assuming they encompass all traffic conditions, real-time decisions and topological features. We acknowledge that the prevailing preferences during the recorded period serve as a guide for feasible routes. The study's noteworthy contribution lies in our departure from analysing individual preferences and trajectory information, instead focusing solely on macro-level measurements of each road segment, such as traffic flow or traffic speed. These types of macro-level measurements are easier to collect compared to individual data sets. We propose an algorithm based on Q-learning, employing traffic measurements within a road network as positive attractive rewards for an agent. In short, observations from macro-level decisions will help us to determine optimal routes between any two points. Preliminary results demonstrate the agent's ability to accurately identify the most feasible routes within a short training period.

2012 ACM Subject Classification Computing methodologies → Q-learning

Keywords and phrases routing, reinforcement learning, q-learning, data mining, macro-level patterns

Digital Object Identifier 10.4230/LIPICs.GIScience.2023.58

Category Short Paper

1 Introduction

The topic of finding routes between two points has been studied in many different fields, such as computer systems, transportation systems and communication networks. The majority of research concentrates on route optimisation, seeking to reduce travel time or distance or to maximise operational efficiencies, such as the maximum number of taxi customers or the maximum storage of a delivery truck. These studies, which employ mathematical optimisation techniques, include optimisation constraints such as the truck's maximum cargo capacity and minimise/maximise the objective function of the main aim, such as travel time. They often take into account the average travel time on a route depending on the length of the road, the timing of the traffic lights, or occasionally the traffic situation, including actual or historical traffic flow and speeds. They also factor in user preferences from surveys or GPS

¹ corresponding author



to generate the most popular routes ahead of time. All of these aspects make optimal route research hard and costly when using multiple data sources. As more realistic findings are sought, models and algorithms become increasingly complex and computationally expensive to investigate every aspect of the traffic situation and road network infrastructure.

In this study, we assume that all of these factors, including traffic user preferences, traffic conditions, and road network features, are already represented in macro-level historical observations. We attempt to extract the most feasible routes using macro-level measurements, in other words, by using the most popular road segments in a road network. We aim to train a reinforcement learning agent to mimic human behaviours for route choices and use the agent to detect the most taken routes between any two nodes which might or might not be optimal routes at that time period under certain traffic conditions.

Although the approach does not guarantee route optimisation, it will identify the most practical and feasible route options at that time from the reflection of the preferences of mass mobility actions. The relevant studies on the algorithms and RL-related studies in route finding in the transport research sector will be briefly discussed in the next section.

2 Related Studies

The principles of routing in transportation are based on the most well-known problems including the travelling salesman problem (TSP), the vehicle routing problem (VRP), and the shortest path problem. They are the primary subjects with the goal of determining the best transport strategies over a road network from a source to a destination. The classic Dijkstra algorithm from 1959 [3] is where the history of discovering shortest paths in networks begins. Heuristic algorithms such as A*[5] concerned with the heading to the destination were created because the Dijkstra algorithm has a vast solution space. These are the core algorithms for static networks, and they only produce one shortest path. However, due to the dynamic nature of road networks, various algorithms for problems involving short paths are introduced with dynamic variables.

Reinforcement learning algorithms have been combined with these conventional techniques to tackle common routing problems such as VRP and TSP [9][14]. Recent RL research has begun to focus on applying deep learning techniques to TSP [13] and VRP problems [1]. These studies are not only limited by classical problems but also attempt to solve optimisation problems in shared transportation [11] and taxi systems [8] by considering future demands. Basically, they define reward systems for desired outcomes such as potential high-demanded areas for taxis. Some studies [2][10] introduce dynamic variables such as energy consumption, and customer request to design some optimisation constraints. This also affects routing studies for passengers [4].

All these studies focus on only the main goal to define a reward system. Our approach will employ mid-rewards to encourage the agent to mimic human behaviours based on observations to find feasible routes. According to the studies [4][7][6], classical routing algorithms are not the best methods to find feasible routes in large systems because of the time complexity and insufficient capabilities of considering only distance costs. On the other hand, All of the aforementioned studies focus on finding the best options within predefined rules, assumptions and constraints. We aim to remove all the pre-defined assumptions and constraints.

■ **Listing 1** Pseudo code for the Q-Learning.

```

Input: Macro-Level measurements (traffic flow), Graph representation
Output: Q-Table
Initialise Reward R(s,a) and Q-table (Q(s,a))
for i:0 to the number of iteration
    Select a random node and its neighbours
    Update the Q-value of the node pairs with the equation
Return Updated Q-table
end
- Select Feasible Route: Reach destinations by selecting the highest
q-values from the starting state
- Derivate other routes: Let the agent choose other best options

```

3 Methodology

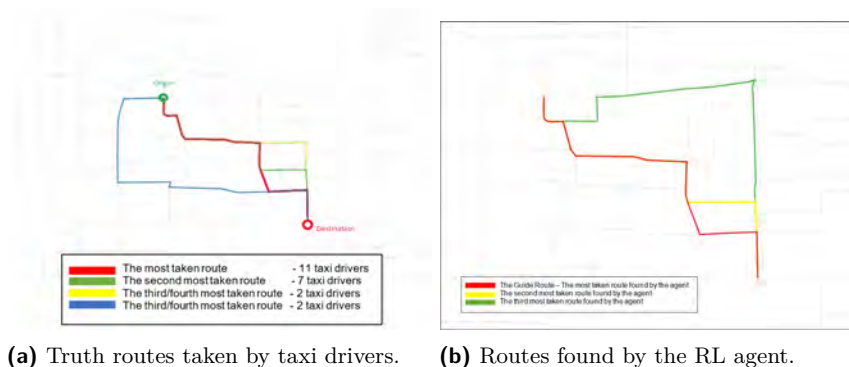
The study's core part is based on a reinforcement learning algorithm called Q-learning. It is a branch of machine learning where an agent maximises its cumulative reward by collecting rewards based on its actions and interactions with an environment. The environment can be modelled mathematically or can be model-free by only focusing on certain rewards that encourage behaviours.

Our approach uses model-free Q-Learning algorithm taking traffic flow values on road segments as the reward. The purpose of the approach is to extract significant routing behaviours from macroscopic observations. It is an offline approach, which implies the agent tries to mimic given behaviours using historical observations without having any impact on the environment. A directed graph is used to represent the environment that represents road networks. Road segments are the edges and intersections are the nodes in this graph. At each intersection, the agent can choose the next road segment (state) to travel through by considering the normalised flow values, in other words, the rewards. These mid-rewards encourage the agent to choose the most taken routes between any two nodes. To help the agent reach the destination point, the highest point is given to the destination points.

These rewards in Q-Learning have an impact on the equation that modifies Q-values in a table (Q-table) displaying the values computed based on state and action pairs. In our approach, we used the Bellmann equation, which performs computations based on states, current and projected rewards, and current Q-values. The Bellman equation;

$$Q_{new}(s_t, a_t) = (1 - \alpha) * Q(s_t, a_t) + \alpha * (R(s, a) + \lambda * \max_{a'} Q(s', a'))$$

Where R is the reward value at state s in the taken action a. The discount variable λ controls the rate at which future rewards will affect the Q-values. The learning rate, or α determines how the current state and actions will influence the Q-values. With the Bellman equation, all the Q-values can be updated in the Q-table in each interaction. This is essential in the agent's training stage. After the training is completed, any starting point can be selected as a current state and the agent can choose the best q-value chain reaching the destination point represented by another state. The highest q-values at each state are consecutively selected to complete this process. The total Q-value value is not required to have the highest value. There might be other route options with higher q-values in total but they are not the best options showing significant mobility patterns. A randomness parameter is introduced to allow the agent to select q-values other than the best one in order



■ **Figure 1** Routes between selected origin-destination pairs.

to derive additional route alternatives. By comparing the distances between the users' actual routes and the routes discovered by the approach, path similarity algorithms can validate the method.

4 Case study

To test the proposed Q-learning approach, a well-known taxi trajectory data set collected in Beijing by Microsoft[12] is used. The OSMnx package, which uses OpenStreetMap as a source, is used to extract the road network. For our approach, there is no need to use micro-level or individual-level measurements such as GPS points or trajectories. However, this data set is helpful to demonstrate the effectiveness of the approach by providing taxi trip trajectories to be used in the validation. We used map-matching algorithms to aggregate all the individual trips for one day in the study area in order to obtain the traffic flow values that the algorithm needs as input. So, we can have the number of taxis at each road segment throughout the entire network.

The dataset contains the trajectory data for nearly 10,000 taxis for the days between February 2nd and February 8th in Beijing. We selected the first-day data of 1000 trips in Beijing's central region for simplicity and due to computation costs. For the study area, there are 657 nodes representing intersections and 1542 edges between these nodes representing each road segment. To determine the origin and destination points, we selected two random point pairs in the concentrated areas having the highest flow values by observing the dataset. Although it is not guaranteed that all trips begin and end at these points, they are sufficient to test our methodology and compare the feasible routes with the real routes passing between these two points. This step is only performed for validation aims.

4.1 Q-learning and Initial Results

During the training phase, the discount and learning rate are chosen as 0.8 and the Q-value table is updated one million iterations. For the study area, this procedure takes 150 seconds to complete.

We detected 31 taxi drivers and 13 different route options taken by these taxi drivers in real life between the origin and destination points in figure 1a. For the visualisation, we showed only the 4 most taken routes by taxi drivers in figure 1a and the three most feasible routes found by the RL agent in figure 1b. The best feasible route extracted from taxi drivers' behaviours by the agent has total Q-values of 6625. In this route, the agent chooses

the best actions at each state until it reaches the destination point. Once the derivation process has been completed, all feasible routes can be extracted by using only traffic flow values, revealing the majority of taxi drivers' choices on their routes. Given the connectivity between nodes, it is clear that there are a finite number of physically feasible routes. The route found by the agent with the best actions is considered the most feasible route and reflects the most seen behaviour of taxi drivers. The other derived options are ordered by the distance of the total q-values from the best feasible option. These q-values can be larger or smaller than the best one in total. All these feasible routes show the preferences of the taxi drivers between these two nodes.

There are two issues with this approach. The agent can choose longer routes to collect more points or it only follows main roads with a high number of traffic flow. Therefore, the reward at the destination point should be decided carefully to encourage the agent to reach the destination point with fewer steps while also avoiding finding only the shortest paths. It should be emphasised that during the training step, all reward values derived from traffic flows are normalised.

The approach demonstrated that the agent can be trained by using only historical macro-level measurements such as traffic flow. These measurements can also be additional traffic state indicators such as traffic speed and travel time. The approach eliminates the requirement for using individual data, which are difficult to obtain and troublesome because of privacy issues. Also, individual preferences and pre-defined assumptions on behaviours can be bypassed by focusing only on macro-level patterns. The approach merely assumes that the best possible sequence of q-values will serve as a guide route.

5 Conclusion

The proposed approach uses a Q-learning-based algorithm to identify feasible route possibilities using just macro-level measurements. It does not have any assumption on the route selection and only mines the historical patterns to detect attracted route segments and extract routes between any two points from observations. This can be seen as a transition to proceed from macro-level measurements to micro-level discoveries. Comparatively speaking, macro-level measurements are simpler and cost-friendly in terms of collecting data than individual data sets like GPS and surveys. They can be received by using sensors, cameras, or even manual counting.

The initial results show that an agent can be trained to extract feasible routes by only exploring the number of vehicles on road segments and sorting the potential options by their selection probabilities. It focuses on the attractiveness and popularity of road segments to take action for the next states. This concept will help us to develop a technique to understand route choice behaviours from macro-level patterns for future research. The approach can combine route set generation and route selection processes in route choice modelling by considering trends at any traffic state. Additionally, these feasible routes may or may not be the shortest ones, the fastest ones, or the ones that taxi drivers select because of the scenic vistas. To uncover the motivations behind these decisions, we will be conducting a more comprehensive analysis.

References


- 1 T Ahamed, B Zou, N P Farazi, and T Tulabandhula. Deep Reinforcement Learning for Crowdsourced Urban Delivery. *Transportation Research Part B: Methodological*, 152:227–257, 2021. doi:10.1016/j.trb.2021.08.015.

- 2 R Basso, B Kulcsár, I Sanchez-Diaz, and X Qu. Dynamic stochastic electric vehicle routing with safe reinforcement learning. *Transportation Research Part E: Logistics and Transportation Review*, 157, 2022. doi:10.1016/j.tre.2021.102496.
- 3 Edsger W. Dijkstra. A note on two problems in connexion with graphs. *Numerische Mathematik*, 1:269–271, 1959.
- 4 Y Geng, E Liu, R Wang, Y Liu, W Rao, S Feng, Z Dong, Z Fu, and Y Chen. Deep Reinforcement Learning Based Dynamic Route Planning for Minimizing Travel Time. In *2021 IEEE International Conference on Communications Workshops, ICC Workshops 2021*, Shanghai, China, 2021. doi:10.1109/ICCWorkshops50388.2021.9473555.
- 5 Peter Hart, Nils Nilsson, and Bertram Raphael. A formal basis for the heuristic determination of minimum cost paths. *IEEE Transactions on Systems Science and Cybernetics*, 4(2):100–107, 1968. doi:10.1109/tssc.1968.300136.
- 6 Y Hu, L Yang, and Y Lou. Path Planning with Q-Learning. In *2021 2nd International Conference on Internet of Things, Artificial Intelligence and Mechanical Automation, IoTAIMA 2021*, volume 1948, North Carolina State University, Raleigh, NC 27695, United States, 2021. IOP Publishing Ltd. doi:10.1088/1742-6596/1948/1/012038.
- 7 F Jamshidi, L Zhang, and F Nezhadalinaei. Autonomous Driving Systems: Developing an Approach based on A* and Double Q-Learning. In *7th International Conference on Web Research, ICWR 2021*, pages 82–85, East China Normal University, Moe International Joint Lab of Trustworthy Software, Shanghai, China, 2021. doi:10.1109/ICWR51868.2021.9443139.
- 8 E Liang, K Wen, W H K Lam, A Sumalee, and R Zhong. An Integrated Reinforcement Learning and Centralized Programming Approach for Online Taxi Dispatching. *IEEE Transactions on Neural Networks and Learning Systems*, 2021. doi:10.1109/TNNLS.2021.3060187.
- 9 T S Mostafa and H Talaat. An Intelligent Geographical Information System for Vehicle Routing (IGIS-VR): A modeling framework. In *13th International IEEE Conference on Intelligent Transportation Systems, ITSC 2010*, pages 801–805, Intelligent Transportation Systems Program, Nile University, 2010. doi:10.1109/ITSC.2010.5625095.
- 10 P Tong, Y Yan, D Wang, and X Qu. Optimal route design of electric transit networks considering travel reliability. *Computer-Aided Civil and Infrastructure Engineering*, 36(10):1229–1248, 2021. doi:10.1111/mice.12678.
- 11 C Wei, Y Wang, X Yan, and C Shao. Look-Ahead Insertion Policy for a Shared-Taxi System Based on Reinforcement Learning. *IEEE Access*, 6:5716–5726, 2017. doi:10.1109/ACCESS.2017.2769666.
- 12 Jing Yuan, Yu Zheng, Xing Xie, and Guangzhong Sun. Driving with knowledge from the physical world. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '11*, pages 316–324, New York, NY, USA, 2011. Association for Computing Machinery. doi:10.1145/2020408.2020462.
- 13 Y Zhang, R Bai, R Qu, C Tu, and J Jin. A deep reinforcement learning based hyper-heuristic for combinatorial optimisation with uncertainties. *European Journal of Operational Research*, 2021. doi:10.1016/j.ejor.2021.10.032.
- 14 M Zolfpour-Arokhlo, A Selamat, S Z Mohd Hashim, and H Afkhami. Modeling of route planning system based on Q value-based dynamic programming with multi-agent reinforcement learning algorithms. *Engineering Applications of Artificial Intelligence*, 29:163–177, 2014. doi:10.1016/j.engappai.2014.01.001.

Moran Eigenvectors-Based Spatial Heterogeneity Analysis for Compositional Data

Zhan Peng ✉ 

Graduate School of Information Sciences, Tohoku University, Sendai, Japan

Ryo Inoue ✉ 

Graduate School of Information Sciences, Tohoku University, Sendai, Japan

Abstract

Spatial analysis of data with compositional structure has gained increasing attention in recent years. However, the spatial heterogeneity of compositional data has not been widely discussed. This study developed a Moran eigenvectors-based spatial heterogeneity analysis framework to investigate the spatially varying relationships between the compositional dependent variable and real-value covariates. The proposed method was applied to municipal-level household income data in Tokyo, Japan in 2018.

2012 ACM Subject Classification Applied computing → Mathematics and statistics

Keywords and phrases Compositional data analysis, Spatial heterogeneity, Moran eigenvectors

Digital Object Identifier 10.4230/LIPIcs.GIScience.2023.59

Category Short Paper

Funding This study was supported by JSPS KAKENHI Grant Number JP21H01447 and JST SPRING Grant Number JPMJSP2114.

1 Introduction

Spatial data that represent parts of a whole and carry only relative information are known as compositional data, such as income structure, land use shares, and vote shares across multiple regions. Although previous studies have considered both the compositional and spatial nature of data [5], little attention has been given to spatial heterogeneity, which is one of the fundamental spatial properties. Spatial heterogeneity in compositional data generally refers to the inconsistent relationships between the relative ratios of each composition and the associated factors across geographical space. This variability can be investigated by estimating spatially varying coefficients (SVCs) at each location [8]. To date, the methodology and application have not been widely discussed.

To enrich this research area, this study proposes a Moran eigenvector-based SVC (MSVC) [3] framework to explore the spatial heterogeneity of compositional data. MSVC links the local variations to the global spatial process, providing interpretable explanations of SVCs. In addition, based on the linear regression framework, MSVC has the advantage of being extendable to accommodate the specific properties of compositional data.

2 Properties of compositional data

Compositional data including D positive components can be represented by a vector $\mathbf{y} = (y_1, \dots, y_D)$, where each component y_j describes only relative information (e.g., proportion or percentage) and all of them sum up to a constant. \mathbf{y} is defined on a simplex space \mathbb{S}^D as



© Zhan Peng and Ryo Inoue;

licensed under Creative Commons License CC-BY 4.0

12th International Conference on Geographic Information Science (GIScience 2023).

Editors: Roger Beecham, Jed A. Long, Dianna Smith, Qunshan Zhao, and Sarah Wise; Article No. 59; pp. 59:1–59:6

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

$$\mathbb{S}^D = \left\{ \mathbf{y} = (y_1, \dots, y_D) \mid y_j > 0, j = 1, \dots, D; \sum_j y_j = 1 \right\}. \quad (1)$$

The constant-sum of compositions leads to spurious correlation [1], which poses challenges to the use of traditional statistical methods with compositional data. A common solution to this problem is to adopt the isometric log-ratio (ILR) transformation [2], which maps compositions \mathbf{y} from the simplex space \mathbb{S}^D to ILR coordinates \mathbf{y}^* in the Euclidean space \mathbb{R}^{D-1} through $\mathbf{y}^* = \text{ilr}(\mathbf{y}) := \mathbf{V}' \ln(\mathbf{y})$. The inverse ILR transformation is $\mathbf{y} = \text{ilr}^{-1}(\mathbf{y}^*) = \mathcal{C} \exp(\mathbf{V}\mathbf{y}^*)$, where \mathcal{C} is the closure operation that $\mathcal{C}\mathbf{y} := \mathbf{y} / \sum_j y_j$. The $D \times (D-1)$ matrix \mathbf{V} obeys $\mathbf{V}' \cdot \mathbf{V} = \mathbf{I}_{D-1}$ and $\mathbf{V} \cdot \mathbf{V}' = \mathbf{I}_D - (1/D)\mathbf{1}_{D \times D}$. Columns \mathbf{v}_i and vectors $\mathbf{e}_i = \mathcal{C} \exp(\mathbf{v}_i)$ forms orthonormal bases of \mathbb{R}^{D-1} and \mathbb{S}^D , respectively. The orthogonality of ILR coordinates allows for the use of classical regression models for each coordinate separately.

3 Method

3.1 MSVC model

The MSVC model is developed based on the correlation between eigenvalues and Moran's I statistic (MC). First, a spatial weight matrix \mathbf{C} is constructed by the binary relationships or distance decaying function (e.g., the exponential function). The eigenvector decomposition $(\mathbf{I} - \mathbf{1}\mathbf{1}'/N)\mathbf{C}(\mathbf{I} - \mathbf{1}\mathbf{1}'/N) = \mathbf{E}_N \mathbf{\Lambda} \mathbf{E}_N'$, where the left-hand side of the equation is also a part of MC, decomposes the spatial structure of the data into a set of orthogonal spatial patterns that are represented by each eigenvector in \mathbf{E}_N . $\mathbf{\Lambda}$ includes the corresponding eigenvalues.

Based on this work, Griffith (2008) [3] introduced a subset of eigenvectors into the basic linear model to account for the spatial heterogeneity in the regressed relationships. The resulting MSVC model is expressed as

$$\mathbf{y} = \sum_{k=0}^K \mathbf{x}_k \circ \beta_k^{ESF} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}). \quad (2)$$

Here, $\beta_k^{ESF} = \beta_k \mathbf{1} + \mathbf{E}\boldsymbol{\gamma}_k$ represents the k -th spatially varying coefficient, which consists of the global trend of the spatial process $\beta_k \mathbf{1}$, and the linear combination of eigenvectors $\mathbf{E}\boldsymbol{\gamma}_k$ that account for the local deviations from the trend at each location. "o" is the column-wise product operator. The next section will extend the MSVC model to accommodate compositional data.

3.2 MSVC model for compositional data

Let $\mathbf{Y} = (\mathbf{y}_1 \ \dots \ \mathbf{y}_N)'$ represent N samples of D -composition data, where \mathbf{y}_i , $i = 1, \dots, N$, is the $D \times 1$ transposed vector of the i -th sample, and $\mathbf{y}_{(j)}$, $j = 1, \dots, D$ is the $N \times 1$ the vector of the j -th component. The ILR transformation of \mathbf{Y} becomes $\text{ilr}(\mathbf{Y}) = (\text{ilr}(\mathbf{y}_1) \ \dots \ \text{ilr}(\mathbf{y}_N))'$, where $\text{ilr}(\mathbf{y}_i) = \mathbf{y}_i^* = (y_{i(1)}^* \ \dots \ y_{i(D-1)}^*)$.

The MSVC model for the j -th ($j = 1, \dots, D-1$) coordinate is formulated as

$$\mathbf{y}_{(j)}^* = \sum_{k=0}^K \mathbf{x}_k \circ \left(\beta_{k(j)}^* \mathbf{1} + \mathbf{E}\boldsymbol{\gamma}_{k(j)}^* \right) + \boldsymbol{\varepsilon}_{(j)}^*, \quad \boldsymbol{\varepsilon}_{(j)}^* \sim \mathcal{N}(\mathbf{0}, \sigma_{(j)}^2 \mathbf{I}). \quad (3)$$

where $*$ denotes the ILR transformation, $\mathbf{x}_k (k = 0, \dots, K, \mathbf{x}_0 = \mathbf{1})$ is the k -th covariate, $\beta_{k(j)}^{SVC*} = \beta_{k(j)}^* \mathbf{1} + \mathbf{E} \gamma_{k(j)}^*$ represents the relationship between the k -th covariate and the j -th ILR coordinate. We can also rewrite the model into a more general form as

$$\mathbf{y}_{(j)}^* = \mathbf{X} \beta_{k(j)}^* + \tilde{\mathbf{E}} \gamma_{k(j)}^* + \boldsymbol{\varepsilon}_{(j)}^*, \tag{4}$$

where $\tilde{\mathbf{E}} = (\mathbf{x}_0 \circ \mathbf{E}, \mathbf{x}_1 \circ \mathbf{E}, \dots, \mathbf{x}_K \circ \mathbf{E})$ are considered as proxy variables. Under the ILR transformation, Equation (4) can be estimated by ordinary linear regression for each $\mathbf{y}_{(j)}^*$, but the interpretation of the estimated coefficients is not straightforward. In line with [4, 8], we adopt the concept of semi-elasticity (SE), which reflects the relative percentage change in a particular composition with respect to a unit change in the covariate of interest. The k -th spatially varying SE of the j -th composition at the i -th location is defined as

$$e(y_j, \mathbf{x}_k)_i = \left(\ln \beta_{ik(j)} - \sum_{m=1}^D y_{i(m)} \ln \beta_{ik(m)} \right) y_{i(j)}. \tag{5}$$

where $y_j, y_{i(m)}, y_{i(j)}$, and $\beta_{ik(j)}$ are the inverse transformed variables in the simplex space.

3.3 Variable selection

Using all eigenvectors can result in an excessive number of explanatory variables. This can create computational challenges and potential overfitting problems. To mitigate these issues, as suggested by [7], we first select eigenvectors whose corresponding eigenvalues satisfy $\lambda_l / \lambda_{max} > 0.25^1$ and then use penalized regression (see Equation (6)) to choose only the eigenvectors that explain significant spatial variations in the data.

$$\min(\mathbf{y}_{(j)}^* - \mathbf{X} \beta_{k(j)}^* - \tilde{\mathbf{E}} \gamma_{k(j)}^*)' (\mathbf{y}_{(j)}^* - \mathbf{X} \beta_{k(j)}^* - \tilde{\mathbf{E}} \gamma_{k(j)}^*) + \lambda |\gamma_{k(j)}^*|_1 \tag{6}$$

The value of λ is determined by cross-validation or information criteria. Because the output of the penalized regression is known to be biased, we use it only for variable selection and apply the proposed model to estimate the coefficients of the selected variables.

4 Empirical application

4.1 Data and methods

We applied the proposed model to the analysis of the municipal-level household income structure of Tokyo, Japan in 2018. The annual income data were aggregated into three main groups: Low (less than 2 million JPY), Middle (between 2 and 7 million JPY), and High (more than 7 million JPY), resulting in a three-composition response variable. The following matrix \mathbf{V} for the ILR transformation of compositions generates two ILR coordinates [6].

$$\mathbf{V} = \begin{bmatrix} 2/\sqrt{6} & 0 \\ 1/\sqrt{6} & 1/\sqrt{2} \\ -1/\sqrt{6} & -1/\sqrt{2} \end{bmatrix}. \tag{7}$$

The first coordinate $\mathbf{y}_{(1)}^*$ refers to the relative importance of the low-income with respect to the other two groups, and the second coordinate $\mathbf{y}_{(2)}^*$ refers to that of the middle-income with respect to the high-income group.

¹ $0.25\lambda_{max}$ relates to roughly 5% of the variance in response variable attributable to positive spatial dependence.

The covariates used in the analysis included the proportion of people with secondary education (Uni), the unemployment rate (Unemp), the proportion of people aged over 65 (Age), and the homeownership rate (House). The data were published by the Statistics Bureau of Japan on the e-Stat portal site (<https://www.e-stat.go.jp/en>). We excluded 11 municipalities with no records, resulting in a final sample size of $N = 51$. Based on the adjacency of regions, we built a spatial weight matrix in which the (i, j) -th element was 1 if two regions i, j shared a common boundary, and 0 otherwise. From this matrix, we extracted 12 out of 51 eigenvectors to be further selected by the penalized regression.

4.2 Results and discussion

First, we conducted the ordinary linear regression without considering the spatial effects. The results shown in Table 1 suggest that all covariates except the unemployment rate are significantly associated with both ILR coordinates. The residual MC indicates that the spatial autocorrelation is significant in $\mathbf{y}_{(1)}^*$, but not significant in $\mathbf{y}_{(2)}^*$.

The results of the proposed model are summarized in Table 2. For $\mathbf{y}_{(1)}^*$, the use of eigenvectors led to a decrease in the residual MC and a noticeable increase in the adjusted R^2 , suggesting that the spatial variations captured by the eigenvectors explain a considerable proportion of the variance in the response variable. No eigenvector was found to be significant on $\mathbf{y}_{(2)}^*$, which aligns with the MC of $\mathbf{y}_{(2)}^*$ shown in Table 1 and proves that the proposed model can distinguish the existence of spatial heterogeneity. This result only indicates that the impacts of covariates on the ratio between middle- and high-income are spatially invariant. However, it does not necessarily imply that the impacts on each income group remain constant. For further analyzing their relationships, we can transform coefficients back to the simplex plane and then calculate the corresponding SEs (Equation (5)).

Figure 1 plots the SEs of each covariate across different income groups. The SEs provide insights into the interconnections among income groups, as they sum up to zero within each region for each covariate. For the entire region, we observe that an increase in the proportion of individuals with secondary education contributes to the shift from low- and middle-income to high-income groups. However, this impact varies by region. Particularly in the southeastern area, which serves as the business and cultural center of Tokyo, the expansion of the high-income group is notably significant. This can be attributed to the concentration of knowledge-intensive industries in this region, which has led to a higher demand for skilled professionals. In Chiyoda-ku, for example, when the proportion of the educated population increases by one unit, the high-income group increases by 0.426%, which

■ **Table 1** Estimation results of the ordinary linear regression.

Variables	$\mathbf{y}_{(1)}^*$		$\mathbf{y}_{(2)}^*$	
	Coefficient	Std. Error	Coefficient	Std. Error
Constant	-0.798*	0.465	0.673**	0.274
Uni	-0.678*	0.381	-1.159***	0.224
Unemp	0.110**	0.049	0.044	0.029
Age	2.967***	1.065	2.988***	0.626
House	-1.460***	0.351	-0.807***	0.206
MC	0.236***		0.014	
Adjusted R^2	0.568		0.860	

Note) : * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

■ **Table 2** Estimation results of the MSVC-based regression.

Variables	$\mathbf{Y}_{(1)}^*$			$\mathbf{Y}_{(2)}^*$		
	Min.	Med.	Max.	Min.	Med.	Max.
Constant	-0.296	-0.239	-0.168		0.673	
Uni	-1.285	-1.048	-0.801		1.159	
Unemp		0.038			0.044	
Age		3.069			2.988	
House	-1.969	-1.757	-1.547		-0.807	
MC		-0.005**			0.014	
Adjusted R^2		0.729			0.860	

Note) : * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

is the highest among all regions. The middle- and low-income groups decrease by 0.153% and 0.273%, respectively. In contrast, in Hinode-machi, which is located on the periphery of Tokyo, the high-income group increases by only 0.189%, and the low- and middle-income groups decrease by only 0.094% and 0.096%, respectively. An increase in the unemployment rate results in the expansion of low- and middle-income groups, along with a decrease in the proportion of the high-income group, primarily observed in southeastern Tokyo. The proportion of people aged over 65 negatively affects the high-income group but positively affects the other two groups. This is consistent with the fact that older people generally have lower incomes and may require more social welfare support. Furthermore, this impact is stronger compared to other factors in terms of the magnitude of the SE, highlighting the importance of considering the impact of the aging of population on income analysis. Lastly, the increase in homeownership rate contributes to the transition of low-income groups into middle-income groups in western and northeastern Tokyo. The middle-income group further shifts to high-income in the southeastern parts.

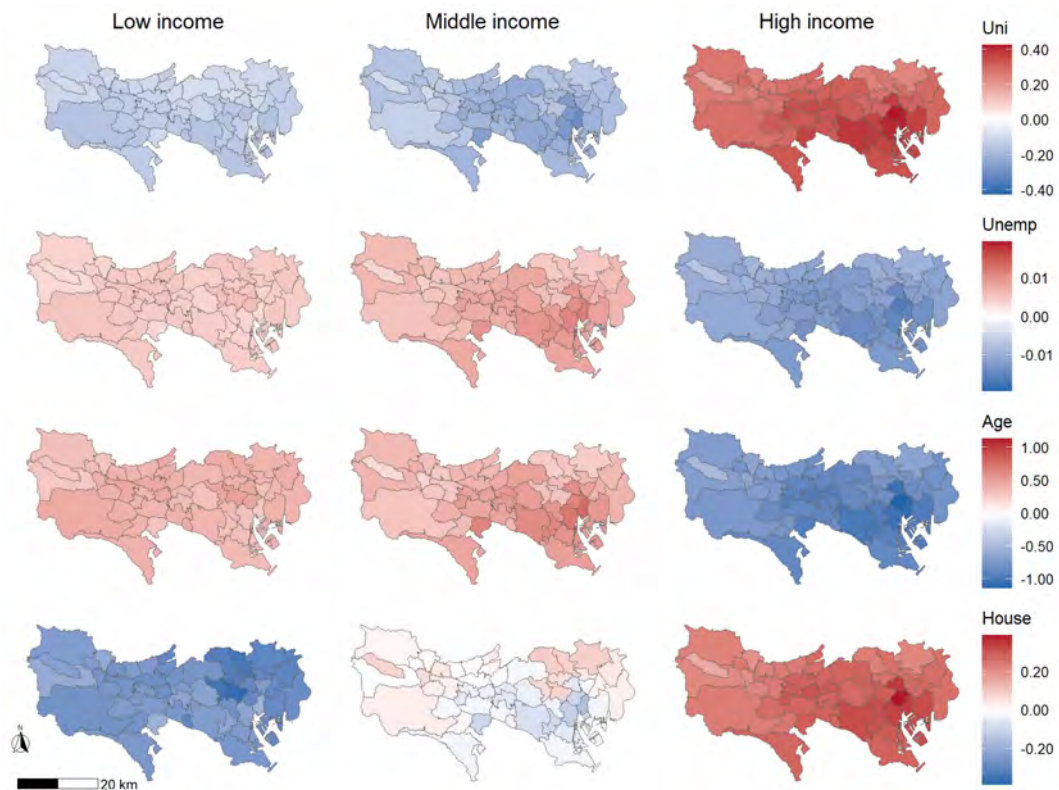
5 Conclusion

This study proposed an MSVC-based framework to investigate the spatial heterogeneity of compositional data. We adopted the ILR transformation and the semi-elasticity to aid the model estimation and interpretation. The application on household income in Tokyo indicated that socio-economic factors affect income distribution differently across regions, which yields insights for understanding the drivers of income inequality.

There are still many challenges and our work is only just beginning. It is worth discussing in the future a more intuitive way of model interpretation. Moreover, an in-depth investigation is necessary to assess the impact a change in the type of spatial weights matrix and the criteria for selecting eigenvectors might have on the outputs. Finally, comparing the performance of the proposed method and previous approaches in analysing spatial heterogeneity would be an interesting topic for future discussions.

References

- 1 J Aitchison. *The Statistical Analysis of Compositional Data*. Chapman & Hall, Ltd., GBR, 1986.
- 2 J. J. Egozcue, V. Pawlowsky-Glahn, G. Mateu-Figueras, and C. Barceló-Vidal. Isometric logratio transformations for compositional data analysis. *Mathematical Geology*, 35(3):279–300, April 2003. doi:10.1023/A:1023818214614.



■ **Figure 1** Spatial distribution of semi-elasticities of MSVC-based CoDA.

- 3 Daniel A Griffith. Spatial-filtering-based contributions to a critique of geographically weighted regression (GWR). *Environment and Planning A: Economy and Space*, 40(11):2751–2769, November 2008. doi:10.1068/a38218.
- 4 Joanna Morais, Christine Thomas-Agnan, and Michel Simioni. Interpretation of explanatory variables impacts in compositional regression models. *Austrian Journal of Statistics*, 47(5):1–25, September 2018. doi:10.17713/ajs.v47i5.718.
- 5 Vera Pawlowsky-Glahn and Juan José Egozcue. Spatial analysis of compositional data: A historical review. *Journal of Geochemical Exploration*, 164:28–32, May 2016. doi:10.1016/j.gexplo.2015.12.010.
- 6 Vera Pawlowsky-Glahn, Juan José Egozcue, and Raimon Tolosana-Delgado. *Modelling and Analysis of Compositional Data*. John Wiley & Sons, 2015. doi:10.1002/9781119003144.
- 7 Hajime Seya, Daisuke Murakami, Morito Tsutsumi, and Yoshiki Yamagata. Application of LASSO to the eigenvector selection problem in eigenvector-based spatial filtering. *Geographical Analysis*, 47(3):284–299, 2015. doi:10.1111/gean.12054.
- 8 Takahiro Yoshida, Daisuke Murakami, Hajime Seya, Narumasa Tsutsumida, and Tomoki Nakaya. Geographically weighted regression for compositional data: An application to the U.S. household income compositions. *GIScience 2021 Short Paper Proceedings. 11th International Conference on Geographic Information Science. September 27-30, 2021*. Poznań:Poland (Online), 2021. doi:10.25436/E2G599.

Toward Causally Aware GIS: Events as Cornerstones

Nina Polous¹   

Institute of Geography and Regional Science, University of Graz, Austria

Abstract

Over the last 50 years, Geographic Information Systems (GIS) have become a vital tool for decision-making. Yet, the increasing volume and complexity of geographical data pose challenges for real-time integration and analysis. To address these, we suggest a causally aware GIS that represents causal relationships. This system uses causality to analyze events and geographical impacts, aiming to offer a more comprehensive understanding of the geographic world. It integrates causality into design and operations, applying robust algorithms and visualization tools for scenario analysis. Unlike traditional GIS, our approach prioritizes an event-based model, emphasizing change as the core concept. This model moves beyond object-oriented models' limitations by considering events as primary entities. The proposed system adopts an event-oriented approach within a Spatio-Temporal Information System, with objects in space and time viewed as event components linked through processes. We introduce an innovative event-based ontology model that enriches GIS by focusing on modeling changes and their interconnections. Lastly, we suggest an IT implementation of this ontology to enhance GIS capabilities further.

2012 ACM Subject Classification Information systems → Geographic information systems

Keywords and phrases Causal Aware GIS, Events, Event-Oriented GIS, Causality

Digital Object Identifier 10.4230/LIPIcs.GIScience.2023.60

Category Short Paper

1 Introduction

GIS has advanced significantly over the past 50 years, transforming geographic research and applications and demonstrating its value to various fields such as urban planning, environmental management, disaster response, through its continuous evolution [11, 12, 8, 21]. GIS has evolved from computer mapping to spatial analysis to solving geographic problems, incorporating our understanding of spatial configurations and perceptions into its approach [25, 24]. Furthermore, GIS is becoming increasingly even more important in our increasingly data-driven world. With its ability to store, manage, and analyze large amounts of geospatial data, GIS aims to provide a powerful tool for solving real-world problems in a variety of domains. This rapid explosion of Geographical data has become one of the biggest challenges facing GIS today. The integration of heterogenous data from multiple sources, making this data available, analyzing it, and using it to make informed decisions is a monstrous task to fulfill. To make this even more challenging, we need to consider in to account that many decisions need to be made real-time or near-real-time in today's increasingly more complex dynamic world. A GIS capable of handling the requirements of a dynamic complex and connected world is yet to be realized. This new GIS not only needs the development of new and innovative methods for visualizing and analyzing geographic data, but more importantly it should be able to represent real-world causal relationships and enable (near)-real-time inference based on continuous flow of data. To address the challenges facing humanity, it is necessary to perform inference of causal relationships, identify effects, and conduct

¹ Corresponding author



© Nina Polous;
licensed under Creative Commons License CC-BY 4.0

12th International Conference on Geographic Information Science (GIScience 2023).

Editors: Roger Beecham, Jed A. Long, Dianna Smith, Qunshan Zhao, and Sarah Wise; Article No. 60; pp. 60:1–60:8

Leibniz International Proceedings in Informatics



LIPIC Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

complex dynamic simulations in human-environment systems [4, 13]. A system that can accurately depict reality's dynamic nature and delineate the relationship between causes and effects, thereby facilitating causal reasoning and inferencing, can aptly be defined as a causality-enabled GIS or, in essence, a causality-aware GIS.

In the next section first the meaning of causality is discussed and the term causal aware GIS and events as its cornerstone are briefly examined. In section 3 differences between event-based models and object-oriented models in GIS are reviewed. In section 4 the proposed event-based model is illustrated and next steps to implement the model in an IT system is listed. Finally, Chapter 5 concludes this article.

2 Causal aware (GIS) systems

Causality refers to the causal relationship between a cause and its resulting effect, where the cause plays a role in producing the effect and the effect is dependent on the cause [5]. "Causality is a relation within the realm of conceptual objects. The relation of cause and effect refers to conceptual events regardless of the relation of the latter to reality" [20]. Causality is a fundamental concept in many fields, including physics, philosophy, psychology, and economics. The earliest recorded inquiry into the relationship between cause and effect can be traced back to Aristotle's *Physics*, which was the first known study of this nature within the realm of science [9]. Philosophy tries "to determine what causal relationships in general are, what it is for one thing to cause another, or what it is for nature to obey causal laws. As I understand it, this is an ontological question, a question about how the world goes on" [23]. Bunge [5] divides the causal problem into two subsets: a) "The ontological problem of causality, i.e. what is causation: what are the characteristics of the causal link; to what extent are such links real; are there causal laws; how do causation and chance intertwine (and so on)?" and b) "The methodological problem of causality, i.e. what are the causation criteria; how do we recognize a causal link and how do we test for a causal hypothesis?" This research focuses on the ontological problem of causality and is interested in causal links, processes as well as causal relations among events.

Causal awareness means a system, agent, or individual's capability to comprehend and account for cause-effect relationships in their environment. This ability enables more than mere correlation, promoting accurate predictions and decisions by considering causal links. Causal awareness aims to enhance decision-making and prediction precision by focusing on underlying mechanisms behind events rather than mere statistical patterns [29]. A causally aware system can thus make informed decisions and accurate predictions by understanding causality, providing a deeper comprehension of complex systems. Developing causally aware geographical systems is key for accurate comprehension of our world, however, systems seamlessly integrating spatiotemporal interactions and causal relationships are still lacking [16]. A causally aware GIS understands and considers cause-effect relationships in a geographic context. Such GIS not only contemplates causal relationships between various events within a geographical space but also leverages causality to analyze interplay between physical, social, environmental, and economic events.

Incorporating causality into GIS involves integrating causal models, algorithms, and data structures for storing and analyzing causal information. To further clarify, causal models refer to the representations that describe the causal mechanisms of a system. These models could be encoded as a system of equations, a directed acyclic graph, or a detailed computational model. Algorithms used in causal analysis commonly include techniques for learning the causal structure from data, methods for causal inference, and procedures for

sensitivity analysis. Data structures suitable for storing and manipulating causal information can include ontologies for representing causal knowledge, database schemas for storing causal data, and file formats for interchange of causal information. These structures ensure efficient retrieval and modification of the causal data. The development of user-friendly interfaces for exploring causal relationships and outcomes is also a critical aspect of integrating causality into GIS. Such interfaces can support users in interpreting the output of causal analyses, navigating through causal structures, and interacting with causal data. This GIS may combine traditional techniques with causality, machine learning, AI, probability, and network analysis methods. A causally aware GIS is a largely unexplored area in GIScience that requires significant research investment. Understanding events, the building blocks of causality [5]. According to Bunge [5] “the causal relation is a relation among events”. Events help us understand causation, the mechanisms linking cause and effect. Events, instances of processes occurring at specific times and places, involve changes in object states. Galton [10] discusses the complexity of causality, including the roles of states, processes, and events, and the challenges of understanding causality from an ontological perspective. Considering these components can enhance understanding of an event and its impact.

Events and their behavioral patterns represent a higher level of knowledge in comparison with changes caused by them. Therefore, they are more valuable for decision makers for making informed decisions. To explore the mechanism of changes, one must investigate the mechanism of events; indeed, events underlie changes [6, 33]. In another word, the Event-based modelling reinforces representation of dynamic behaviors of geographical phenomena, generation of hypothesis, investigation of scientific complex relationships, and ability to explore causal relationships among associated entities while providing an opportunity to understand underlying procedures [3]. In this research “event” as the basic units of causality is further explored and discussed in the next section. While the current paper provides a preliminary outline for a causality-aware GIS centered on events, it is important to note that it is impossible to cover the breadth and complexity of the literature on event ontology in this writing. However, to gain a more in-depth understanding of the connection between processes and events, readers are referred to the vast body of work by Antony Galton or references such as [31] and [1].

3 From object-oriented view toward event-based models in GIS

Initially, GIS modeled geographical features independent of time due to their long-lasting identities and locations [29]. However, in the late 80s and early 90s, GIS started to incorporate time, addressing geographic feature dynamics [2, 19]. This allowed for recording object history and predicting future changes. Still, the focus remained on geographical features, with time stamps tracking feature states [32]. This object change view, reflecting ontologies that have dominated western thought since Aristotle’s time [30], sees the world as a collection of classified objects with specific properties, relationships, and behaviors. Hägerstrand [15] highlights the importance of time in human activities to assess the dynamic behaviour of people in space, especially the motion of individuals in space and time. Miller [26] and Yuan [34] have exhibited this fact in their work on transportation and urban analysis, and on analysis of physical phenomena, such as storms. Different researchers such as Miller [26] have promoted the work of Hägerstrand’s under the principal of geo-spatial lifelines. However, Hornsby and Egenhofer [17] deal with the object change view through the concept of identity-based change. There are several downsides when modelling changes with the object change view [19, 32]: first, expensive computations and calculations are needed to

detect and identify changes between snapshots. Second, developing or imposing rules for internal reasoning is challenging, since there is no understanding of the restrictions upon the temporal structure. Third, no matter what the size of changes is, a full snapshot is produced at each time sequence leading to storing huge amount of redundant information. Fourth, when models concentrate on objects' changes rather than a snapshot, it becomes challenging to identify "when and what change becomes so substantial that an object is no longer the same object". Due to such restrictions in the object-oriented model, many researchers have suggested event-based models as an alternative solution [7, 28, 33, 32, 29].

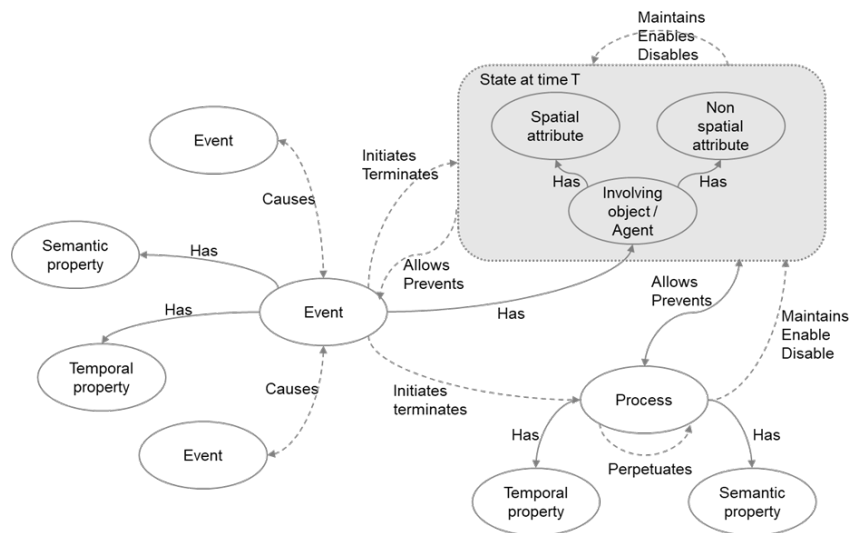
In event-based models, change is the main concept that is modelled and change units are the primary items for analysis and evaluation. Claramunt and Thériault [7] define events as things which occur. Particularly they explain that processes cause changes in the state of objects, these changes reveal the outcome of the process and create events. The event perspective sees objects in space and time merely as information elements of the events, which are connected to other event elements through internal or external processes [29]. Peuquet [27] defines an event as indicator of changes in a place or an object. Peuquet and Duan [28] refer to an event as a way to represent spatiotemporal manifestation of processes. Worboys [32] and Worboys and Hornsby [33] define an event as a happening that should be differentiated from a thing or continuant. They suggest that events are necessary to record the mechanism of change. Events are perhaps the most extensive information container for dynamic geo-historical phenomena and geographical reality [29]. To explain any event well enough, we should take into account its objective and results, its individual participants, its place in space and time, and its relationships to various other events. Representing enough large number of events along these dimensions may enable us to analyze and discover underlying social historical processes of the globe [14]. Although early calls to maintain and preserve records of events and processes to understand dynamic behaviours of the reality go back to late 80s [6], its realization in GI Systems is far from being called done [29].

Most early GIS data models can be considered as expansions of cartographic models, and existing methods to organise and store data generally use data layers and spatial blocks [18]. This radically limits expressing reach relationships between geographic elements at different scales, the mechanisms of interaction among elements, and their evolutionary processes with different semantic meanings and multiple attributes [22]. Hence, the need to move the concept of "representing geographical reality" beyond the principle of mapping objects which have distinct spatial, temporal and attributive identities as usual in object-oriented systems [29]. Although GIS is an information system, its core idea is to explore the geographical reality and real-world complexities, its patterns, processes, and reach interactions among different geographical phenomena, to enable us to understand the world better. Using a generic event-oriented perspective to implicitly represent causal relationships among different components of a Spatio-Temporal Information System makes realization of this goal possible. Thus, the core of GIS should follow the mission to explore the laws of nature and reveal its essence to humanity, which cannot be achieved by considering events and process as second-class elements in today's GIS systems. Leveraging event-oriented representations of reality, enables GIS to serve as a true knowledge representor of real-world complexities and move toward a causal-aware system.

4 Event-based models

Spatiotemporal ontologies have been extensively researched, but there is a notable gap in explicitly considering events as entities within GIS. Previous studies focused on modeling events and their relationships, often treating the temporal dimension as an attribute of

spatial objects or as a part of new entities called “spatiotemporal objects.” To bridge this gap, this study introduces a conceptual model to mirror and manage real-world dynamism. In this novel system, events unfold, processes occur, and states change. The study goes beyond conventional mapping practices that view objects as static geographic entities. Instead, it centers on modeling change as the core concept, with the analysis and evaluation primarily based on change units. This approach prioritizes the temporal dimension, recognizing the crucial role of recording event sequences over time. To handle the complex relationships between spatial and temporal dimensions, new methods are necessary. Events serve as identity containers for objects, states, and processes, forming the fundamental components for mapping dynamic phenomena. By treating events as first-class objects in GIS, this proposed event-oriented perspective enables the modeling of causality and complex relationships in our dynamic world. The mapping perspective shifts towards viewing the world as a network of interconnected relationships, unlocking richer language and understanding interactions among objects, events, and underlying processes.



■ **Figure 1** Schematic concept (modified version of model proposed by Polous [29]).

Figure 1 illustrates the schematic concept of the event-centric perspective. This new perspective integrates two aforementioned mapping principles; an event centric approaches that look at the phenomena holistically while in its turn inherited the object-oriented perspective for mapping the object in reductionist way. In this new integrated model, objects belong to states while processes are running on them and making changes in their states (spatial and none-spatial) through the power of events as causal forces. In fact, the states and processes together create a new concept so called ‘dynamic snapshot’ at each moment. The snapshots contain both processes and objects, therefore they are no longer static but have an inherent dynamism which provides a solid foundation for understanding events which are happening over time. The events can initiate or terminate a state through initiating or terminating an external or internal process. Indeed, by looking at the dynamic snapshots we can see different objects, in various states which are undergoing particular processes. The snapshots are constantly renewed as time passes; the snapshots alter from one moment to the next, because the present elements in the snapshot are changing. Here, the events are considered as fixed historical records so as time passes, event are occurring, and getting gradually added, numbered and stored in the event database. This new perspective offers a Spatio-temporal Information System a standard way to mathematically model the changing world while developing a firm basis for the logical modelling of dynamical systems.

In the pursuit of incorporating an event-based ontology model within an IT system, the Author is implementing a meticulous seven-step strategy, utilizing key tools like the Web Ontology Language (OWL) and the Resource Description Framework (RDF). The first step revolves around defining the ontology using an OWL ontology tool such as Protégé, including elements such as events, processes, states, and involved objects. OWL allows the creation of detailed and consistent models by providing greater machine interpretability than XML, RDF, and RDFS. Its reasoning capabilities enable the automation of data consistency verification and allow querying beyond instance retrieval. Subsequently, instance data tailored for particular use cases will be populated within the ontology. As a third step, we will integrate the defined ontology with other existing GIS models. The fourth stage involves data storage through RDF, facilitated by an RDF database. RDF is a standard model for data interchange, offering broad interoperability, which enables the integration of data from various sources. Its graph-based data model provides flexibility in representing knowledge, allowing users to structure and link data in any way. The data will then be queried through the SPARQL query language in the fifth stage, unveiling hidden connections and relationships within the data. SPARQL, with its capabilities to express queries across diverse data sources, supports complex reasoning tasks and extraction of valuable insights from the semantic data.

These first five steps are enough to conduct needed research for any specific use-case, however, to expand the reach of the system, the Authors aim to make the IT System available to others through APIs and User interfaces. The sixth step focuses on the development of APIs or web services, integrating the knowledge management system with additional GIS applications and allowing for external querying and support in decision-making processes. To conclude, a user-friendly interface and visualization tools will be constructed to foster user interaction with the ontology, thus improving the overall usability of the system. By applying this model to an IT system underpinned by Semantic Web technologies, we anticipate constructing a robust knowledge management system. This system will empower users to navigate intricate relationships, inform decision-making processes, and provide valuable insights for a myriad of stakeholders. Through the leverage of Web Semantic languages such as OWL and RDF, this model offers extensive manipulation and reasoning capabilities, enabling users to create, manage, exchange, and reason with knowledge about resources. This expands the range of capabilities and empowers users to generate and explore complex relationships and hypotheses. The findings from this endeavor will be published in due course.

5 Conclusion

GIS strives to offer a complete and accurate understanding of geographic data, but the explosion of data complexity and the dynamic nature of our world poses challenges. Thus, the next logical step is to develop a causally aware GIS - a system that understands and integrates causal relationships and supports real-time decision making. A causally aware GIS enhances data analysis by considering causal relationships between various factors in a geographical context. To realize this system, we need to infuse causality into its design, operations, and analysis processes. This includes integrating causal models, algorithms, and structures that support the manipulation and analysis of causal information. Crucially, the system should be capable of computing various scenarios' effects and outcomes and clearly representing causal information. This needs to be complemented with user-friendly visualization tools for exploring causal relationships and their implications. Historically, GIS models had limitations in expressing the interaction and evolution between geographic

elements. An event-based model, which sees change as the primary concept being modeled, can better represent dynamic geo-historical phenomena. As a next step, this paper proposes the development of a causally aware GIS system that comprehensively represents reality and understands causal relationships. This requires innovative methods for visualizing and analyzing geographic data, coupled with a deep grasp of causality. Implementing the proposed event-based ontology model within an IT system is a pivotal step in this direction, involving seven systematic steps from constructing the ontology to developing user-friendly interfaces and visualization tools.

References

- 1 João Almeida, Ricardo Falbo, and Giancarlo Guizzardi. *Events as Entities in Ontology-Driven Conceptual Modeling*, pages 469–483. SpringerLink, October 2019. doi:10.1007/978-3-030-33223-5_39.
- 2 Marc P Armstrong. Temporality in spatial databases. In *GIS/LIS 88 Proceedings: Accessing the World*, pages 880–889, 1988.
- 3 Kate Beard. Modelling change in space and time: An event-based approach. In *Dynamic and Mobile GIS; Investigating Changes in Space and Time*, pages 55–76. CRC Press, 2006.
- 4 Susanne Bleisch, Matt Duckham, Antony Galton, Patrick Laube, and Jarod Lyon. Mining candidate causal relationships in movement patterns. *International Journal of Geographical Information Science*, 28(2):363–382, 2014. doi:10.1080/13658816.2013.841167.
- 5 Mario Bunge. *Causality and Modern Science: Third Revised Edition*. Dover Publications, 2009.
- 6 N. R. Chrisman. *Beyond the snapshot: changing the approach to change, error, and process*, pages 85–93. Oxford University Press, New York, NY, USA, 1998.
- 7 Christophe Claramunt and Marius Thériault. Managing time in gis an event-oriented approach. In James Clifford and Alexander Tuzhilin, editors, *Recent Advances in Temporal Databases*, pages 23–42, London, 1995. Springer London.
- 8 Max Egenhofer, Keith Clarke, Song Gao, Teriitutea Quesnot, Randolph Franklin, May Yuan, and David Coleman. *Contributions of GIScience over the Past Twenty Years*, pages 9–34. Advancing Geographic Information Science: The Past and Next Twenty Years. GSDI Association Press, February 2015.
- 9 A Falcon. The stanford encyclopedia of philosophy. *Encyclopedia*, 2015.
- 10 Antony Galton. States, processes and events, and the ontology of causal relations. *Frontiers in Artificial Intelligence and Applications*, 239:279–292, January 2012. doi:10.3233/978-1-61499-084-0-279.
- 11 Michael F Goodchild. Twenty years of progress: GIScience in 2010. *Journal of Spatial Information Science*, 2010(1), July 2010. doi:10.5311/josis.2010.1.2.
- 12 Michael F. Goodchild. Reimagining the history of GIS. *Annals of GIS*, 24(1):1–8, January 2018. doi:10.1080/19475683.2018.1424737.
- 13 Michael F. Goodchild and J. Alan Glennon. Representation and computation of geographic dynamics. In *Understanding Dynamics of Geographic Domains*. CRC Press, 2008.
- 14 Karl E. Grossner. Event objects for spatial history. In *Extended Abstracts Volume, GIScience 2010, Zurich*, 2010.
- 15 Torsten Hägerstrand. WHAT ABOUT PEOPLE IN REGIONAL SCIENCE? *Papers in Regional Science*, 24(1):7–24, January 1970. doi:10.1111/j.1435-5597.1970.tb01464.x.
- 16 Yufeng He, Yehua Sheng, Barbara Hofer, Yi Huang, and Jiarui Qin. Processes and events in the centre: a dynamic data model for representing spatial change. *International Journal of Digital Earth*, 15(1):276–295, 2022. doi:10.1080/17538947.2021.2025275.
- 17 Kathleen Hornsby and Max J. Egenhofer. Identity-based change: a foundation for spatio-temporal knowledge representation. *International Journal of Geographical Information Science*, 14(3):207–224, 2000. doi:10.1080/136588100240813.

- 18 Chris B. Jones. *Geographical Information Systems and Computer Cartography*. Routledge, May 2013. doi:10.4324/9781315846231.
- 19 Gail Langran. *Time in Geographic Information Systems*. CRC Press, November 1992. doi:10.1201/9781003062592.
- 20 Victor Fritz Lenzen. *Causality in Natural Science*. Springfield, Ill., Thomas, 1954.
- 21 Paul A. Longley. *Geographic Information Science and Systems, 4th Edition*. Wiley, New York, 2015.
- 22 Guonian Lü, Michael Batty, Josef Strobl, Hui Lin, A-Xing Zhu, and Min Chen. Reflections and speculations on the progress in geographic information systems (gis): a geographic perspective. *International Journal of Geographical Information Science*, 33(2):346–367, 2019. doi:10.1080/13658816.2018.1533136.
- 23 John Leslie Mackie. *The Cement of the Universe: A Study of Causation*. Oxford, England: Oxford, Clarendon Press, 1974.
- 24 David Mark. Geographic information science: Critical issues in an emerging cross-disciplinary research domain. *Journal of the Urban and Regional Information Systems Association*, 12:45–54, January 2000.
- 25 David M. Mark, Christian Freksa, Stephen C. Hirtle, Robert Lloyd, and Barbara Tversky. Cognitive models of geographical space. *International Journal of Geographical Information Science*, 13(8):747–774, 1999. doi:10.1080/136588199241003.
- 26 Harvey J Miller. What about people in geographic information science? *Computers, Environment and Urban Systems*, 27(5):447–453, 2003. doi:10.1016/S0198-9715(03)00059-0.
- 27 Donna J. Peuquet. It’s about time: A conceptual framework for the representation of temporal dynamics in geographic information systems. *Annals of the Association of American Geographers*, 84(3):441–461, 1994. URL: <http://www.jstor.org/stable/2563777>.
- 28 Donna J. Peuquet and Niu Duan. An event-based spatiotemporal data model (estdm) for temporal analysis of geographical data. *International Journal of Geographical Information Systems*, 9(1):7–24, 1995. doi:10.1080/02693799508902022.
- 29 Nina Polous. *Event Cartography: A New Perspective in Mapping*. Dr. Hut Verlag, Munich, 2016.
- 30 N Rescher. The stanford encyclopedia of philosophy. *Encyclopedia*, 2008.
- 31 Fabrício Henrique Rodrigues, Mara Abel, Valerio Basile, Tommaso Caselli, and Daniele P. Radicioni. What to consider about events: A survey on the ontology of occurrents. *Appl. Ontol.*, 14(4):343–378, January 2019. doi:10.3233/A0-190217.
- 32 Michael Worboys. Event-oriented approaches to geographic phenomena. *International Journal of Geographical Information Science*, 19(1):1–28, 2005. doi:10.1080/13658810412331280167.
- 33 Michael Worboys and Kathleen Hornsby. From objects to events: GEM, the geospatial event model. In *Geographic Information Science*, pages 327–343. Springer Berlin Heidelberg, 2004. doi:10.1007/978-3-540-30231-5_22.
- 34 May Yuan. Representing complex geographic phenomena in gis. *Cartography and Geographic Information Science*, 28(2):83–96, 2001. doi:10.1559/152304001782173718.

Mobility Vitality: Assessing Neighborhood Similarity Through Transportation Patterns In New York City

Dan Qiang   

Platinal Analysis Lab, Department of Geography, McGill University, Montréal, Canada

Grant McKenzie   

Platinal Analysis Lab, Department of Geography, McGill University, Montréal, Canada

Abstract

Though numerous studies have examined human mobility within an urban environment, few have explored the concept of urban vitality purely through the lens of urban transportation. Given the importance of different modes of transportation within a city, such analysis is necessary. In this short paper, we introduce the novel concept of mobility vitality by integrating human mobility and urban vitality, offering a multilayered framework to assess the degree of transportation and mobility within and between regions. The mobility patterns of three transportation modes, namely subway, taxicab, and bike-share, are first examined independently. These patterns are then aggregated to form the composite measure of static mobility vitality. Through this measure, we evaluate similarities between neighborhoods. Our results observed significant spatial differences in the travel patterns of three transportation modes on weekdays and weekends. Moreover, neighborhoods with high static mobility vitality have relatively similar mobility patterns. Ultimately, this approach aims to find neighborhoods with imbalanced transportation infrastructure or inadequate public.

2012 ACM Subject Classification Information systems → Geographic information systems; Applied computing → Transportation

Keywords and phrases mobility vitality, mobility similarity, transportation, bike-sharing, taxi, subway, New York City

Digital Object Identifier 10.4230/LIPIcs.GIScience.2023.61

Category Short Paper

1 Introduction

In 1961, the urban activist Jane Jacobs introduced the concept of *urban vitality* as a qualitative measure of a city's pulse [2]. The idea suggests that varying tempos of human activities and pedestrian flow can all be employed to differentiate regions. For decades, most of the research related to this concept was done using qualitative surveys, demographic studies, and narrative analysis. The difficulties with such approaches are costly and labor-intensive and are prone to subjective biases. The recent dramatic growth of publicly accessible activity and mobility data has set the stage for alternative approaches to assessing urban vitality.

Despite a large body of literature targetting the extraction of individual human mobility patterns and their accompanying impact variables [1, 7], little attention has been paid to urban dynamics characterized purely by individual movement. Recently, a growing number of research teams have focused on temporal characteristics of mobility to better understand urban vitality [3]. For instance, Sulis et al. [6] examined smart-card rail trips to assess spatiotemporal variation in urban vitality in London. They produced a set of three dynamic properties, namely the number of people, the continuity, and the fluctuations of this presence over particular intervals of time. Similarly, Zeng et al. [9] created a new index to measure urban vitality based on records from a bicycle-sharing system. Further work has demonstrated



© Dan Qiang and Grant McKenzie;
licensed under Creative Commons License CC-BY 4.0

12th International Conference on Geographic Information Science (GIScience 2023).

Editors: Roger Beecham, Jed A. Long, Dianna Smith, Qunshan Zhao, and Sarah Wise; Article No. 61; pp. 61:1–61:6
Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

that *lively* regions of a city correlate with taxi drop-off locations [10]. A variety of research has shown that urban vitality/vibrancy can be measured through data ranging from social check-ins and points of interest to trajectories and mobile phone data [4, 8].

Though progress is being made, research focused exclusively on mobility as a measure of urban vitality is lacking [11]. In exploring the vitality of different parts of a city through a mobility lens, one is able to identify the impact that access to different mobility modes, has on city cohesion. Furthermore, a combination of mobility signatures can be used as a measure through which different regions of a city can be compared [5]. Urban and transportation planners can use such a measure to better understand the impacts of policy decisions on the vibrancy and vitality of the city as a whole. Through integrating human mobility with urban vitality, we proposed the novel concept of *mobility vitality*, serving as a multilayered framework to evaluate the degree of transportation and mobility within a space. In this preliminary work, we aim to address the following two research questions (RQ).

- RQ1 Can a region, *e.g.*, neighborhood, be quantified by the mobility patterns of different modes of transportation that exist and traverse the region? Furthermore, do these patterns vary by mode and region?
- RQ2 Can mobility vitality, as represented by a combination of mobility patterns, be used to compare and differentiate regions within the same city?

We address these questions through an analysis of three different modes of transportation within New York City (NYC). As the most densely populated city in the United States, NYC's transportation ecosystem is both complex and extensive. The scale of our analysis is neighborhoods within the five boroughs of NYC and the extent of analysis varies based on the service area of each transportation system.

2 Data and Analysis

To start, three data sets representing three very different modes of transportation were collected. These include bike-share, subway (rail), and taxicab data. We restricted our analysis to May 2019, cleaned the data to remove errors, and aggregated the month of data to days in a typical week. We use this week as a representative sample of transportation usage in NYC. May was chosen due to the limited holidays, historically decent weather, and fewer people on summer vacation. For micro-mobility, we accessed data for the widely used bicycle-sharing system, *Citi Bike*¹. Citi Bike is a privately operated docking station-based bike-sharing system. Citi Bike trip data include the start and end times of each trip as well as the origin and destination stations. For mid-sized transportation, we accessed trip data for *yellow taxis*². The yellow taxi trip records include fields capturing pick-up and drop-off dates, times, and locations. For mass transit, we analyzed turnstile data of the *NYC subway system*³. These data report an accumulated number of entrances and exits, per station at a four-hour temporal resolution. All data were cleaned to remove erroneous trips (e.g., those that were one minute in length, 200 miles, etc).

Next, we intersected the trip data with the NYC neighborhood boundaries⁴ to assign trip volume for each of the three services to each neighborhood in NYC. The assigned volume includes both origins (entries) and destinations (exits). More specifically, the numbers of

¹ <https://citibikenyc.com/system-data>

² <https://data.cityofnewyork.us/Transportation/2019-Yellow-Taxi-Trip-Data/2upf-qytp>

³ <https://data.ny.gov/Transportation/Turnstile-Usage-Data-2019/xfn5-qji9>

⁴ <https://data.cityofnewyork.us/City-Government/2020-Neighborhood-Tabulation-Areas-NTAs-Tabular/9nt8-h7nd>

origins and destinations were combined to determine the final trip volume. For the subway turnstile, the total number of entries and exits for every turnstile within a station was summed. For example, there are four control areas in the “Cortlandt St.” station and each control area has 10 turnstiles. The trip volume for that station was calculated as the sum of all passengers through the 40 turnstiles. The trip data were then divided by the populations of their respective neighborhoods. This process was straightforward for the bike-share and subway turnstile data as they are represented as point geometries. The taxicab trip data, however, is reported by polygonal taxi zone⁵ (TZ). A dasymetric mapping approach was used to allocate taxicab trip origins and destination TZ data to the NYC neighborhood boundaries.

To address RQ1, our static⁶ *mobility vitality* measure was generated by summing the individual transportation mobility patterns across each region, producing a single value for each neighborhood. While we took an “equal weights” approach here, the measure is designed to allow a user to adjust the importance (weights) of each individual transportation mode in the overall mobility vitality result, depending on their interests. Given this measure of mobility vitality, we then examined how such a measure could be used to better understand the vitality and variability of mobility services within a city such as NYC. To start, we averaged the mobility vitality measure for each neighborhood by weekday and weekend. This allowed us to subtract weekend mobility vitality from weekdays to better identify temporal variations in mobility and differentiate neighborhoods based on prototypical commuting behavior. Finally, we examined mobility vitality as a measure on which to identify similarities between neighborhoods based purely on how inhabitants and visitors use different transportation systems. To address this RQ2, we used Jensen-Shannon divergence (JSD), a method for assessing the (dis)similarity of two probability distributions. In our case, we took the trip volume for each day of the week of our three transportation modes as a *distribution*. Having one distribution for each neighborhood allowed us to assess the similarity between all neighborhood pairs. We then identified the neighborhoods that were most similar to all other neighborhoods and those that were most unique.

3 Results and Discussion

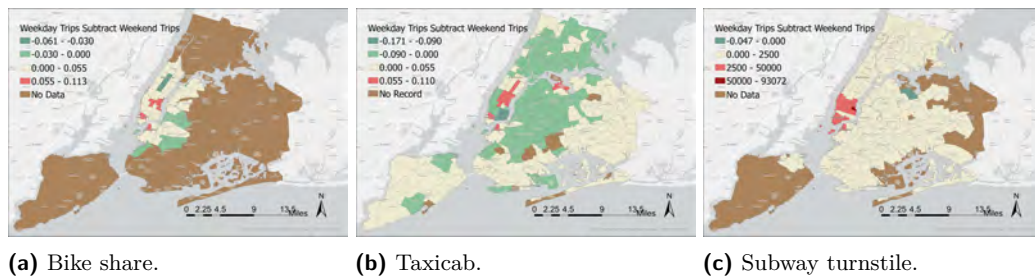
For all three modes of transportation, there is greater mobility activity on weekdays than on weekends. For bike-share origins and destination points, the population-normalized mean values are 0.028 and 0.021, for weekdays and weekends, respectively. Similarly, the mean population-normalized taxi pick-up density on weekdays is 0.0143, while on weekends it is 0.0137. The subway turnstile data was much more pronounced with a population-normalized weekdays value of 1,407.86 and a weekend value of 839.91. These large values indicate that, for many of the neighborhoods within NYC, the number of subway passengers is several orders of magnitude higher than the residential population.

The weekday/weekend variation in normalized transportation trips is shown in Figure 1. In order to compare weekday trips and weekend trips, we delineated the legend on the maps by setting 0 as the dividing line in the class intervals. In both bike and taxi categories, those values greater than 0 and those less than 0 were separately averaged into two intervals. For the metro map, given the significantly higher number of weekday trips compared to weekend ones, only one level was established for values less than 0, while those greater

⁵ <https://data.cityofnewyork.us/Transportation/NYC-Taxi-Zones/d3c5-ddgc>

⁶ Static here refers to the fact that temporal variability was not included in this approach.

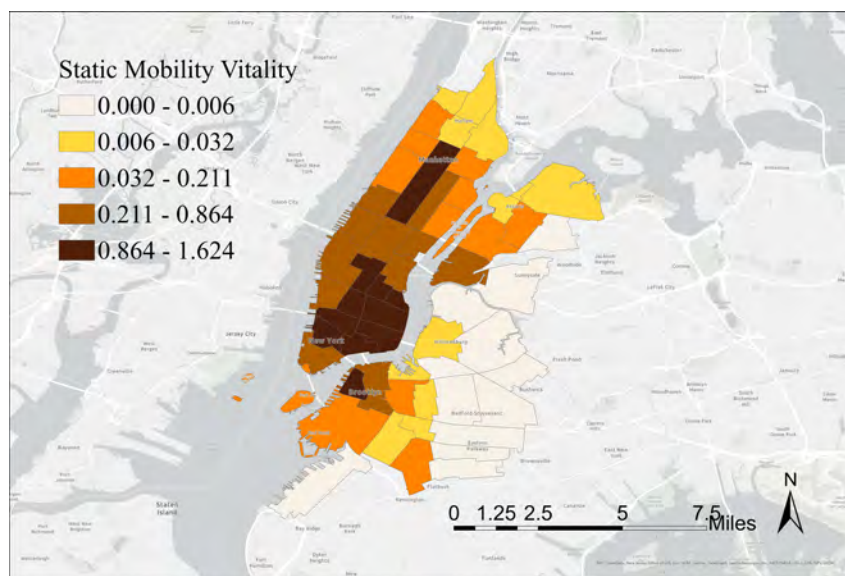
61:4 Mobility Vitality



■ **Figure 1** Population-normalized weekend trip counts subtract from weekday trip counts, for three modes of transportation.

than 0 were evenly divided into three levels. In general, the Manhattan business district witnesses a predominance of weekday trips over weekend ones, with the intensity varying across transportation modes. Bike sharing predominantly favors weekdays, with Central Park being the exception. Conversely, neighborhoods encompassing recreational areas report higher weekend bike trip volumes. Taxi trips exhibit a starkly distinct pattern, with higher weekend volumes in both northern and southern Manhattan, notably in downtown neighborhoods near Queens. Subway data, however, shows a universal weekday preference, except in East Elmhurst and North Corona. A clear spatial clustering of neighborhoods with the greatest discrepancy between weekday and weekend trips is evident in downtown Manhattan.

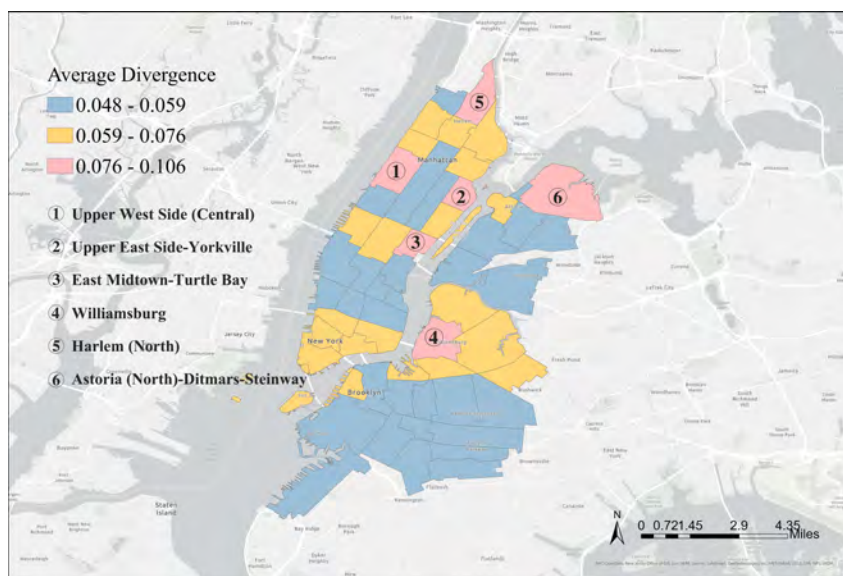
The results of the equally-weighted static mobility vitality measure are shown in Figure 2. The operating region for the bike share service is the most spatially restrictive of our data and so all data sets were restricted to this analysis area. As one can see, the greatest degree of mobility vitality is in Central Park and the southeast corner of Manhattan. As one moves towards the east side of Brooklyn and north Harlem, the vitality gradually decreases.



■ **Figure 2** Static mobility vitality as calculated by summing the population-normalized trip volume from three different modes of transportation.

The results of our Jensen-Shannon divergence approach are shown in Figure 3. In this Figure, pink neighborhoods are the most unique neighborhoods in terms of mobility vitality, reporting the highest average JSD values. These include the Upper West Side (Central),

Upper East Side-Yorkville, East Midtown-Turtle Bay, Williamsburg, Harlem (North), and Astoria (North)-Ditmars-Steinway. Among these, four are located in Manhattan, while Queens and Brooklyn each contain one. Comparing these results to the static mobility vitality map shown in Figure 2, it can be observed that the six most dissimilar neighborhoods are not the ones with the highest static mobility vitality. They belong to the lower-scoring group in terms of the three individual mobility patterns as well as static mobility vitality. This speaks to the influence of the temporal dimension on assessing the vitality of a city with respect to mobility. In our data, most neighborhoods with high static mobility vitality have relatively low divergence values, indicating that they tend to be similar to one another. Neighborhoods with low JSD values are in regions with high volumes of everyday traffic for each mode of transportation and their individual mobility patterns show little variation. The six most unique neighborhoods, as measured by our mobility patterns, are scattered throughout the city but share a common characteristic, they are all waterfront neighborhoods.



■ **Figure 3** Unique and similar neighborhoods as measured through three modes of transportation using Jensen-Shannon divergence.

4 Conclusions and Next Steps

In this preliminary work, the concept of mobility vitality is proposed to measure the degree of transportation and mobility within a region. This work investigates mobility vitality patterns when different transportation starts or ends in the neighborhood and uses these patterns to identify the divergence between different neighborhoods within NYC. Not surprisingly, we found that mobility patterns are different on weekdays than on weekends. In most cases, the volume of trips in downtown neighborhoods is greater during weekdays than on weekends; however, taxicabs in some central business districts are the exception. Additionally, neighborhoods with excessive divergence are dispersed and more dissimilar neighborhoods often exhibit a high degree of clustering and high mobility vitality.

The next steps for this work will involve including additional modes of transit and assessing the robustness of our approach through varying types of transportation. Our current mobility vitality approach is meant as a “proof-of-concept” and further iterations will

allow users to vary the weights depending on the question they are investigating. Last, the current analysis was conducted using data collected at a daily temporal resolution. We aim to examine the spatiotemporal characteristics of mobility vitality with a finer time granularity in the future.

The results of the analysis presented in this short paper are meant to offer a glimpse at the objective of generating a mobility vitality measure that represents the spatial and temporal dynamics of mobility within a city. Through developing such a measure, our aim is to empower urban and transportation planners with measures by which similarities and differences within a city can be identified. Planners and government agencies will be able to monitor how transportation policies can change vitality within a city and use such a measure to improve equitable access to transportation systems within the urban environment.

References

- 1 Ezgi Eren and Volkan Emre Uz. A review on bike-sharing: The factors affecting bike-sharing demand. *Sustainable Cities and Society*, 54:101882, 2020. doi:10.1016/j.scs.2019.101882.
- 2 Jane Jacobs. *The death and life of great American cities*. Random House, New York, 1961.
- 3 Chaogui Kang, Dongwan Fan, and Hongzan Jiao. Validating activity, time, and space diversity as essential components of urban vitality. *Environment and Planning B: Urban Analytics and City Science*, 48:1180–1197, 2021. doi:10.1177/2399808320919771.
- 4 Qian Li, Caihui Cui, Feng Liu, Qirui Wu, Yadi Run, and Zhigang Han. Multidimensional Urban Vitality on Streets: Spatial Patterns and Influence Factor Identification Using Multisource Urban Data. *ISPRS International Journal of Geo-Information*, 11(1):2, 2022. doi:10.3390/ijgi11010002.
- 5 Grant McKenzie and Daniel Romm. Measuring urban regional similarity through mobility signatures. *Computers, Environment and Urban Systems*, 89:101684, 2021. doi:10.1016/j.compenvurbsys.2021.101684.
- 6 Patrizia Sulis, Ed Manley, Chen Zhong, and Michael Batty. Using mobility data as proxy for measuring urban vitality. *Journal of Spatial Information Science*, 16:137–162, 2018. doi:10.5311/JOSIS.2018.16.384.
- 7 Yang Xu, Dachi Chen, Xiaohu Zhang, Wei Tu, Yuanyang Chen, Yu Shen, and Carlo Ratti. Unravel the landscape and pulses of cycling activities from a dockless bike-sharing system. *Computers, Environment and Urban Systems*, 75:184–203, 2019. doi:10.1016/j.compenvurbsys.2019.02.002.
- 8 Yihong Yuan and Martin Raubal. Analyzing the distribution of human activity space from mobile phone usage: an individual and urban-oriented study. *International Journal of Geographical Information Science*, 30(8):1594–1621, 2016. doi:10.1080/13658816.2016.1143555.
- 9 Peng Zeng, Ming Wei, and Xiaoyang Liu. Investigating the Spatiotemporal Dynamics of Urban Vitality Using Bicycle-Sharing Data. *Sustainability*, 106(1):1714, 2020. doi:10.3390/su12051714.
- 10 Bin Zhang, Shuyan Chen, Yongfeng Ma, Tiezhu Li, and Kun Tang. Analysis on spatiotemporal urban mobility based on online car-hailing data. *Journal of Transport Geography*, 82:102568, 2020. doi:10.1016/j.jtrangeo.2019.102568.
- 11 Zhonghao Zhang, Yusi Zhang, Tian He, and Rui Xiao. Urban Vitality and its Influencing Factors: Comparative Analysis Based on Taxi Trajectory Data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15:5102–5114, 2022. doi:10.1109/JSTARS.2022.3183176.

An Evaluation of the Impact of Ignition Location Uncertainty on Forest Fire Ignition Prediction Using Bayesian Logistic Regression

David Röbl  

Institute of Geodesy, Graz University of Technology, Graz, Austria

Rizwan Bulbul  

Institute of Geodesy, Graz University of Technology, Graz, Austria

Johannes Scholz   

Institute of Geodesy, Graz University of Technology, Graz, Austria

Mortimer M. Müller  

Institute of Silviculture, University of Natural Resources and Life Sciences, Vienna, Austria

Harald Vacik  

Institute of Silviculture, University of Natural Resources and Life Sciences, Vienna, Austria

Abstract

This study investigates the impact of location uncertainty on the predictive performance of Bayesian Logistic Regression (BLR) for forest fire ignition prediction in Austria. Historical forest fire ignitions are used to create a dataset for training models with the capability to assess the general forest fire ignition susceptibility. Each recorded fire ignition contains a timestamp, the estimated location of the ignition and a radius defining the area within which the unknown true location of the ignition point is located. As the values of the predictive features are calculated based on the assumed location, and not the unknown true location, the training data is biased due to input uncertainties. This study is set to assess the impact of input data uncertainty on the predictive performance of the model. For this we use a data binning approach that splits the input data into groups based on their location uncertainty and use them later for training multiple BLR models. The predictive performance of the models is then compared based on their accuracy, area under the receiver operating characteristic curve (AUC) scores and brier scores. The study revealed that higher location uncertainty leads to decreased accuracy and AUC score, accompanied by an increase in the brier score, while demonstrating that the BLR model trained on a smaller high-quality dataset outperforms the model trained on the full dataset, despite its smaller size. The study's contribution is to provide insights into the practical implications of location uncertainty on the quality of forest fire susceptibility predictions, with potential implications for forest risk management and forest fire documentation.

2012 ACM Subject Classification Theory of computation → Bayesian analysis

Keywords and phrases Forest Fire Prediction, Ignition Location Uncertainty, Bayesian Logistic Regression, Bayesian Inference, Probabilistic Programming

Digital Object Identifier 10.4230/LIPIcs.GIScience.2023.62

Category Short Paper

Funding The research was carried out as part of the IGNITE – Improving the Assessment of Forest Fire Susceptibility project, which is funded by the Austrian Forest Fund (Federal Ministry of Agriculture, Forestry, Environment and Water Management).



© David Röbl, Rizwan Bulbul, Johannes Scholz, Mortimer M. Müller, and Harald Vacik; licensed under Creative Commons License CC-BY 4.0

12th International Conference on Geographic Information Science (GIScience 2023).

Editors: Roger Beecham, Jed A. Long, Dianna Smith, Qunshan Zhao, and Sarah Wise; Article No. 62; pp. 62:1–62:7

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Introduction

The impact of forest fires in Europe has been increasingly severe due to climate change, leading to longer fire seasons, expansion of affected areas, and unprecedented conditions for fire-fighting services [12]. In countries such as Austria, forest fire prediction models, which form the backbone of early warning systems, use manually collected incident reports to predict the outbreak and behaviour of forest fires. However, uncertainty in the input data, due to human involvement makes the data susceptible to various uncertainties. In order to create reliable predictive models for forest fires, it is essential to understand how input uncertainty impacts the accuracy of predictions. This study specifically investigates the impact of uncertainty surrounding the initial fire ignition point location on the accuracy of forest fire ignition predictions. Bayesian Logistic Regression (BLR) is a flexible approach for predictive modeling, particularly with input data uncertainties. It provides a robust mathematical model to quantify uncertainty, incorporate prior knowledge, and improve the model's generalization. Unlike the point estimates provided by traditional Logistic Regression (LR), the Bayesian method provides a full predictive posterior distributions, that quantifies input data and model uncertainty [4]. The primary objective of this study is to analyze the sensitivity of BLR models to forest fire ignition location uncertainty by training multiple models using training datasets with different levels of associated uncertainty. For this purpose this study utilizes the Austrian forest fire database, which stores the locations of past fire ignition points. Each point is associated with a positional uncertainty in the form of a distance radius, which determines the area where the forest fire may have started, as shown in Figure 1. The paper is organized as follows. In section 2 we elaborate on the related work, followed by the methodology described in section 3. This section covers data preparation, model training and evaluation. Section 4 covers the results and section 5 discusses the results achieved.



Figure 1 This map displays recorded fire ignition locations and their associated buffers, indicating the uncertainty of each ignition position. The slope raster underneath provides further insight into the terrain, showcasing strong variations within the uncertainty regions.

2 Related Work

Logistic Regression (LR) has been used extensively in wildfire science and management, according to [10], who provided a comprehensive review of Machine Learning (ML) applications in this area. BLR, on the other hand, has seen limited use in wildfire prediction. [5] applied BLR with uninformed priors to estimate the probability of large fires based on weather

components, while [8] trained hierarchical BLR models with different priors to estimate the probability of fire occurrence based on forest vulnerability and climatic conditions. While previous studies have investigated the impact of weather conditions, land cover, and human activities on the predictive performance of wildfire fire ignition models using LR and other complex ML methods, few have examined the effect of location uncertainty on predictive models. [1] conducted a study to analyze the impact of fire ignition location uncertainty on kernel density estimates by systematically displacing ignition points and comparing the resulting density surfaces. In their study on wildfire prediction in Portugal, [7] utilized LR models. They found that the recorded ignition locations used for model training had a margin of error of up to 500 meters. However, they argued that the impact of this positional error on predictions could be considered negligible due to the large sample size and the small scale of the geospatial data used in their study. To the best of our knowledge, no study has yet investigated the impact of location uncertainty on the predictive performance using BLR models.

3 Methodology

3.1 Data Sources

In this study, the primary data source used was the Austrian forest fire database, which was established within the activities of European and nationally funded projects (AFRI and ALP FIIRS) [13]. This database covers forest fire incidents beginning in the 16th century with an almost complete documentation of forest fires events since the beginning of the 21st century and provides valuable information such as the coordinates of the assumed ignition point location, the location uncertainty radius, the cause of the fire, and the size of the affected area. The scope of this study was limited to human-caused fire incidents that occurred between 2001 and 2018 and have a location uncertainty of no more than 500 meters. A total of 955 fire events were considered in the analysis. To generate predictive features we used additional data sources covering a digital elevation model (data.gv.at; 10x10m), a building and population raster (100x100m), the street network (gip.gv.at) and a vegetation type raster (bfw.gv.at; 10x10m). All data layers were projected to the Austria Lambert reference system.

3.2 Data Preparation

To get an evenly balanced data set, we randomly sampled 1085 points within the forest domain, which we used as non-fire events. The study encompasses several features, namely: distance to buildings, population density, distance to roads, road type, distance to bicycle and pedestrian pathways, vegetation type, elevation, slope and aspect. These specific features were chosen, drawing upon the research conducted by [2] and [3]. The values associated with these features are calculated based on the incident point location. Finally, the recorded fire incidents are divided into four groups based on their associated distance radius, representing the uncertainty of the fire ignition location. The first group, serving as the validation set, includes all samples with an uncertainty smaller or equal to 100 meters. The other three groups, serving as training datasets, are created based on uncertainty thresholds that ensures a roughly equal distribution of samples across the groups. Furthermore, the training data from all groups are combined into a single additional training set. Table 1 provides an overview of the four groups and their corresponding uncertainty ranges and sample sizes.

■ **Table 1** Overview of training and validation data groups.

Bin	Uncertainty Range (meter)	Size	Distribution (non-fire, fire)
Validation	[0, 100]	429	228, 201
Training 1	(100, 250]	580	319, 261
Training 2	(250, 400]	527	275, 252
Training 3	(400, 500]	504	263, 241
Training full	(100, 500]	1611	857, 754

3.3 Model Training

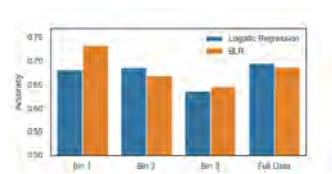
Each training dataset is used to fit both a traditional LR and a BLR model. LR is a statistical method that is well-suited for modeling binary outcomes, such as the presence or absence of forest fires. Unlike traditional LR, BLR assigns a prior probability distribution to the regression coefficients, which reflects prior beliefs about the relationship between the features and the outcome. By using Bayesian inference, the prior is combined with the likelihood of the observed data to obtain the posterior probability distribution of the coefficients. Our choice of prior distribution was a Student-T distribution with a mean of 0, a scale of 2.5, and 1 degree of freedom, resulting in a Cauchy distribution. This prior distribution is known to allow for robust inference and has been recommended for weakly informative priors in Bayesian analysis [9]. Before fitting the data to the model parameters, the numerical input features were standardized to improve model convergence. We utilized scikit-learn (scikit-learn.org) for traditional LR and the probabilistic programming library PyMC (pymc.io) for BLR, which leverages the Markov Chain Monte Carlo (MCMC) algorithm for Bayesian inference.

3.4 Model Evaluation

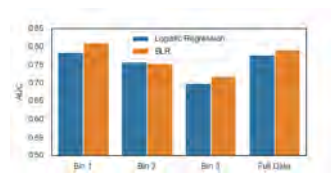
To assess the predictive performance of the various models on the validation set, we employ two common metrics: accuracy and area under the receiver operating characteristic curve (AUC). Accuracy is defined as the proportion of correctly classified incidents (i.e., whether a fire occurred or not) based on a threshold of 0.5 for the predicted probability values. AUC, on the other hand, measures the ability of the model to distinguish between fire and non-fire cases across all possible threshold values. Both accuracy and AUC have a scale from 0 to 1, where values above 0.5 suggest performance that exceeds random guessing. When evaluating the danger of forest fires, it's important to consider the probability values provided by the model, rather than just the binary classification. These values represent the model's uncertainty in identifying potential fires and indicate the danger of a fire starting under the observed conditions. Therefore, we additionally assess the quality of the probability estimates using the brier score. The brier score measures the average difference between the predicted probability and the actual outcome. A higher score indicates that the model's probability estimates are less reliable, while a lower score indicates greater reliability. The brier score ranges from 0 to 1 and was first introduced in [6].

4 Results

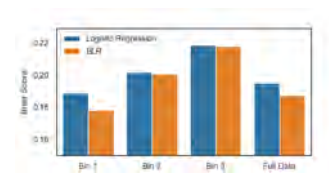
The reported accuracy, AUC and brier scores for the BLR models are mean values of 10 runs. Since the variation among the different outcomes is low, we do not report all model runs in this short paper. Figure 2 and Figure 3 depict accuracy and AUC scores for the LR and BLR models trained on the different datasets. The results clearly show that the



■ **Figure 2** Accuracy (with threshold = 0.5) of LR and BLR models.



■ **Figure 3** AUC score of LR and BLR models.



■ **Figure 4** Brier score of LR and BLR models.

model performance decreases with increasing ignition location uncertainty. For the BLR models, there is a +8,6% accuracy, a +9,3% AUC and a -4% brier score (as shown in Figure 4) difference between the model trained on the high quality dataset (100-250 meter location uncertainty) and the model trained on the poor quality dataset (400-500 meter). The BLR trained on the high quality dataset even outperformed the BLR model trained on the full dataset (+4,6% accuracy, +1,8% auc and -1% brier score). When comparing the BLR and LR models, it can be seen that the BLR model trained on the high quality dataset performs significantly better than the LR model trained on the same data (+5,2% accuracy, +2,5% AUC). However, this observation does not apply to the models trained on the other datasets, except for the brier score (Figure 4), where BLR consistently outperforms LR by a small margin.

5 Discussion

The findings of this study highlight the impact of location uncertainty on the predictive performance of fire ignition models. The bias resulting from uncertainty about the true location of the fire ignition has a significant effect on the models' accuracy, with a clear decrease in performance as the location uncertainty increased in the training data. This phenomenon is attributed to location bias affecting all spatial features, especially those with high spatial variability, such as slope. Given the relatively small number of data samples available for forest fire ignitions in Austria, a critical question arises about whether using high-quality data (in terms of location uncertainty) is more advantageous than employing all available data with mixed quality for training purposes. Our study indicates that BLR is a suitable method for dealing with small data sets. It achieves better results when trained on a small high-quality dataset than when trained on a mixed-quality dataset containing roughly three times as many samples. In contrast, the traditional LR model trained on the high-quality data only achieves similar results as the one trained on the full dataset. The reason behind this is, that BLR allows prior knowledge to be incorporated regarding the relationship between the predictors and outcome variable. This incorporation works as a regularizer, constraining overfitting or underfitting in small datasets by reducing the parameter estimates towards the prior distribution. However, an extensive analysis of different prior distributions in our BLR model was not conducted, neglecting the fact that different features may require different sets of priors. Furthermore, there is an additional point that requires discussion. The interpretation of the probability values generated by the forest fire ignition prediction models can be somewhat ambiguous. While the probability score can be an indicator of the level of danger, it can also be viewed as a measure of uncertainty in the model's prediction. However, [11] argue that these two concepts, the predicted level of danger and the prediction uncertainty, should be treated separately. This suggests the need

to investigate how we can use Bayesian inference, which provides additional information about the prediction uncertainty, to communicate both the predicted probability and the model's uncertainty to decision-makers in forest fire management.

6 Conclusion

In summary, this study highlights the importance of considering location uncertainty in fire ignition models, and the potential benefits of using BLR for dealing with small datasets. The findings of this study can have significant implications for forest fire management and documentation, as they suggest that investing in a high-quality dataset and utilizing BLR with weakly informed priors may help overcome the limitations posed by a small training dataset.

References

- 1 Giuseppe Amatulli, Fernando Pérez-Cabello, and Juan de la Riva. Mapping lightning/human-caused wildfires occurrence under ignition point location uncertainty. *Ecological modelling*, 200(3-4):321–333, 2007.
- 2 Natalie Arndt, Harald Vacik, Valerie Koch, Alexander Arpaci, and Hartnut Gossow. Modeling human-caused forest fire ignition for assessing forest fire danger in Austria. *iForest-Biogeosciences and Forestry*, 6(6):315, 2013.
- 3 Alexander Arpaci, Bodo Malowerschnig, Oliver Sass, and Harald Vacik. Using multi variate data mining techniques for estimating fire susceptibility of tyrolean forests. *Applied Geography*, 53:258–270, 2014.
- 4 Umang Bhatt, Javier Antorán, Yunfeng Zhang, Q Vera Liao, Prasanna Sattigeri, Riccardo Fogliato, Gabrielle Melançon, Ranganath Krishnan, Jason Stanley, Omesh Tickoo, and others. Uncertainty as a form of transparency: Measuring, communicating, and using uncertainty. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 401–413, 2021.
- 5 Ross A Bradstock, JS Cohn, A Malcolm Gill, Michael Bedward, and C Lucas. Prediction of the probability of large fires in the Sydney region of south-eastern Australia using fire weather. *International Journal of Wildland Fire*, 18(8):932–943, 2009.
- 6 Glenn W Brier and others. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3, 1950.
- 7 Filipe X Catry, Francisco C Rego, Fernando L Bação, and Francisco Moreira. Modeling and mapping wildfire ignition risk in Portugal. *International Journal of Wildland Fire*, 18(8):921–931, 2009.
- 8 Georgios Charizanos and Haydar Demirhan. Bayesian prediction of wildfire event probability using normalized difference vegetation index data from an Australian forest. *Ecological Informatics*, 73:101899, 2023.
- 9 Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, 3rd edition, 2013. doi:10.1201/b16018.
- 10 Piyush Jain, Sean C P Coogan, Sriram Ganapathi Subramanian, Mark Crowley, Steve Taylor, and Mike D Flannigan. A review of machine learning applications in wildfire science and management. *Environmental Reviews*, 28(4):478–505, 2020.
- 11 Meelis Kull and Peter A Flach. Reliability maps: a tool to enhance probability estimates and improve classification accuracy. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2014, Nancy, France, September 15-19, 2014. Proceedings, Part II 14*, pages 18–33. Springer, 2014.

- 12 J. San-Miguel-Ayanz, T. Durrant, R. Boca, P. Maianti, G. Libertá, T. Artés-Vivancos, D. Oom, A. Branco, D. de Rigo, D. Ferrari, H. Pfeiffer, R. Grecchi, M. Onida, and P. Löffler. Forest Fires in Europe, Middle East and North Africa 2021. Technical report, Publications Office of the European Union, Luxembourg, 2022. doi:10.2760/34094.
- 13 Harald Vacik, Natalie Arndt, Alexander Arpaci, Valerie Koch, Mortimer Mueller, and Hartmut Gossow. Characterisation of forest fires in Austria. *Austrian Journal of Forest Science*, 128(1):1–31, 2011.

Calculating Shadows with U-Nets for Urban Environments

Dominik Rothschedl ✉ 🏠

dwh GmbH, Vienna, Austria

Franz Welscher ✉ 

Institute of Geodesy, Graz University of Technology, Austria

Franziska Hübl ✉ 

Institute of Geodesy, Graz University of Technology, Austria

Ivan Majic ✉ 

Institute of Geodesy, Graz University of Technology, Austria

Daniele Giannandrea ✉ 🏠

dwh GmbH, Vienna, Austria

Institute of Information Systems Engineering, TU Vienna, Austria

Matthias Wastian ✉ 🏠

dwh GmbH, Vienna, Austria

Johannes Scholz ✉ 

Institute of Geodesy, Graz University of Technology, Austria

Niki Popper ✉ 🏠

dwh GmbH, Vienna, Austria

Institute of Information Systems Engineering, TU Vienna, Austria

Abstract

Shadow calculation is an important prerequisite for many urban and environmental analyses such as the assessment of solar energy potential. We propose a neural net approach that can be trained with 3D geographical information and predict the presence and depth of shadows. We adapt a U-Net algorithm traditionally used in biomedical image segmentation and train it on sections of Styria, Austria. Our two-step approach first predicts binary existence of shadows and then estimates the depth of shadows as well. Our results on the case study of Styria, Austria show that the proposed approach can predict in both models shadows with over 80% accuracy which is satisfactory for real-world applications, but still leaves room for improvement.

2012 ACM Subject Classification Computing methodologies → Neural networks

Keywords and phrases Neural Net, U-Net, Residual Net, Shadow Calculation

Digital Object Identifier 10.4230/LIPIcs.GIScience.2023.63

Category Short Paper

Funding The presented results were obtained within the project PV4EAG (888491) funded by the Austrian Research Promotion Agency (FFG) <https://www.ffg.at/>.

1 Introduction

The production of renewable energy in urban environments is a crucial contribution to carbon neutrality. This requires the assessment of the solar energy potential that is reflected by the solar radiation on the earth's surface [1]. Of particular interest is the assessment of solar energy potential in urban environments, where almost 50% of the world's population is located. Besides photovoltaic systems mounted on roofs, there is additional potential for



© Dominik Rothschedl, Franz Welscher, Franziska Hübl, Ivan Majic, Daniele Giannandrea, Matthias Wastian, Johannes Scholz, and Niki Popper;
licensed under Creative Commons License CC-BY 4.0

12th International Conference on Geographic Information Science (GIScience 2023).

Editors: Roger Beecham, Jed A. Long, Dianna Smith, Qunshan Zhao, and Sarah Wise; Article No. 63; pp. 63:1–63:6

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

photovoltaic systems on facades. As urban areas are covered by buildings that cast shadows on surrounding buildings, and the production of renewable energy with photovoltaics is influenced by shadows - the calculation of shadows is key to make informed decisions.

Contemporary Geographic Information Systems (GIS) are capable of representing shadows for 3D city models based on some type of surface information. The task of generating shadows is usually performed by strictly geometrical approaches such as GIS shadow calculation models. Such models have high accuracy, but usually come at high computational costs depending on spatial and temporal resolution of the data and calculation [11, 12]. One such model is used as a source of ground truth in this study as well [4, 5].

In order to address these issues, this paper presents an approach to calculate shadows using GeoAI methods. One approach, already using machine-learning libraries (tensor-based techniques) but simply optimizing the data preparation and computation time, was shown by [2]. Their urban test area is also represented by a digital surface model (DSM) with a spatial resolution of $1m$. They provide a proof-of-concept for binary shadow calculation, in contrast to our ML-approach, which is able to predict not only the binary value, but also the depth of the shadow. In detail, we present a method for the calculation of the shadow depth of tiles in Styria using U-Net [13]. With this machine-learning approach to solving this problem of physics, we strive to get results more quickly after a computationally intensive training [16, 17].

The U-Net, as basic structure of our network, was originally developed for segmenting biomedical images and is designed to get by with few training images and to be able to localise high-resolution features. These properties fit well for our shadow segmentation task, because as the shadow calculation depends on the position of the sun, we would have needed a large amount of training data.

2 Methodology

The architecture of a U-Net is made up by a contracting path followed by an expansion path, which is roughly symmetric to the contraction. Each contraction step consists of two convolutions and a subsequent max pooling as well as a doubling of the channel numbers. In every expansion step we have an upsampling, followed by a concatenation with the channels of the same size from the contracting path and two convolutions. The concatenations between contraction and expansion paths are key to allow for better localization of high-resolution features and thus more precise segmentation. On top of the U-Net we include residual connections [8] within the convolution net, concretely we insert identity mappings between every other multi-channel feature map. This further eases the training of our net, allowing us to use a larger number of layers in our net.

For our purposes, a U-Net with 5 up- and downsampling layers and a depth of 256 in the bottleneck layer was implemented, where the input tensor consists of the geographic data for each pixel in a 64×64 tile. More specifically, for each pixel, the input is defined as the elevation information, i.e. *surface height* and the *surface height plus the average height of the objects* in the pixel, the terrain information *slope* and *aspect*, and *sun angle* and *azimuth* at a certain time. The input tensor therefore has the size $64 \times 64 \times 6$. Let's take as an example a tile whose centre has the coordinates (47.0867407955596, 15.423575649486619). The pixel in the upper right corner of the tile has a surface height of 352.3 metres, with the objects on the surface 354.59 metres. The terrain has an aspect of 0.23° and a slope of 15.77° . Assuming that the shadow is to be calculated on 19.02.2022 at 12 noon, the angle of the sun is 39.28° and the azimuth is 240.971° . Hence the input tensor of this pixel is (352.3, 354.59, 15.77, 0.23, 39.28, 240.97).

The output differs between the two models. The binary model calculates whether a pixel is shaded or not, whereas the shadow depth model tries to predict the shadow depth, or more precisely the degree of shading of objects with consideration of the shadow depth in a pixel. The degree of shading is divided into eleven classes, where the first class indicates that 0% of the object surface is shaded, the second class indicates that 10% of the object surface is shaded, the third class indicates that 20% of the object surface is shaded, and so on. This is therefore a multi-class net.

For the training of the networks, training areas in Styria (Austria) were defined and divided into 64×64 tiles, where one pixel corresponds to $1m^2$. While the elevation and terrain information for the input layer were derived from the digital surface model, a random day of the year 2022 at 12 o'clock was chosen for each tile to determine the position of the sun, from which the azimuth and the angle of the sun were calculated. A total of 449152 tiles were generated and further augmented, i.e., rotated 90, 180 and 270 degrees to increase the size of training data. This dataset was split into 66% training tiles, whereas the rest serve as validation tiles. For each of these tiles, the ground truth was calculated to train the nets. For this purpose, a QGIS Terrain Shading plugin was used to calculate the shadow depth over the DSM [5]. The next step is to transform the result of the QGIS plugin to make it comparable to the output of the nets. For the binary model, a pixel is not shaded, i.e. it has the value 0, if the shadow depth is zero, otherwise its value is 1. For the multi-class net, the degree of shading of an object p_{shaded} , if there is any shading, is calculated from the depth of shading $d_{shadow} \in \mathbb{R}^-$ and the object height, the difference between surface height $h_{surface} \in \mathbb{R}^+$ and ground level $h_{ground} \in \mathbb{R}^+$.

$$p_{shaded} = \begin{cases} 0 & , d_{shadow} = 0 \text{ and } h_{surface} - h_{ground} = 0 \\ 1 & , d_{shadow} < 0 \text{ and } h_{surface} - h_{ground} = 0 \\ \lfloor \frac{d_{shadow}}{h_{surface} - h_{ground}} \rfloor & , \text{ else} \end{cases} \quad (1)$$

Hence, $p_{shadow} \in \{0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1\}$ which is equivalent to eleven categories.

In each training step, different evaluation metrics were applied to the current results to check whether the neural net is learning. One of those is the Jaccard index [6], also known as intersection over union, and the other one is the Dice score [7]

$$\mathcal{J}(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}, \quad (2)$$

$$\mathcal{C}(A, B) = \frac{2|A \cap B|}{|A| + |B|}, \quad (3)$$

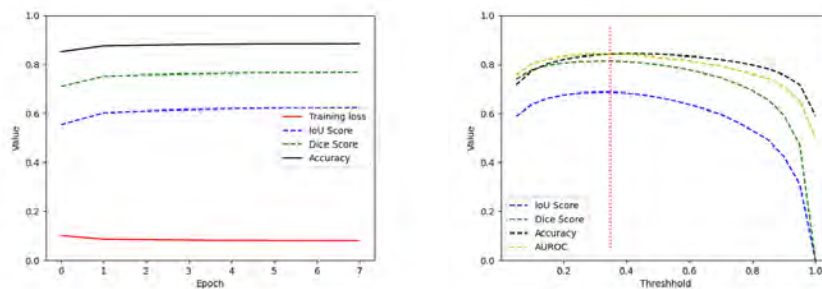
where A and B are any two batches of tiles to be compared. Both scores determine the similarities of sets and are common [10, 15].

3 Results

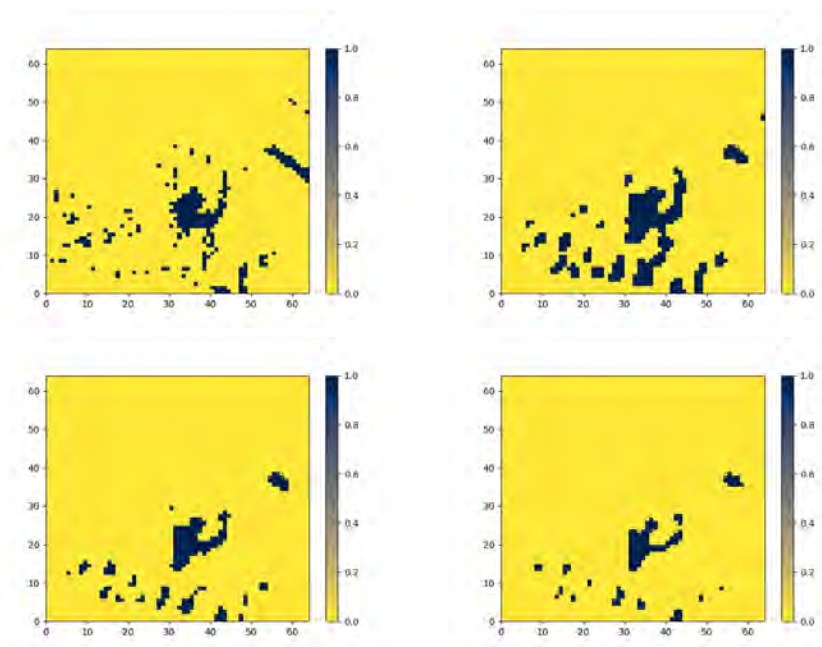
3.1 Binary Model

As already discussed in Section 2 the binary model predicts whether a pixel is shadowed or not. To measure the learning behaviour of the network, we will use the Jaccard index (IoU), Dice Score and Accuracy, and then perform a threshold analysis.

63:4 Calculating Shadows with U-Nets for Urban Environments



■ **Figure 1** The left figure shows the training metrics of the binary model. Due to early stop algorithm, training was cancelled after 8 epochs. The right figure shows the IoU, Dice, Accuracy and AUROC [3] for validation dataset at different thresholds, where the vertical line indicates the threshold with the best results.

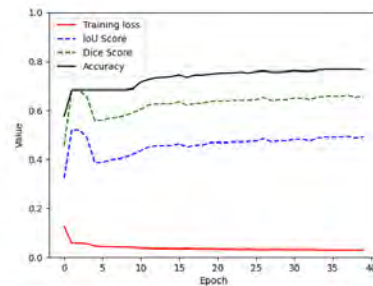


■ **Figure 2** The top left figure shows **the original shadowing** computed with the QGIS plugin. The top right figure shows the **shading predicted** by the binary model **with threshold 0.2**. The bottom left figure shows the **shading predicted** by the binary model **with threshold 0.35**. The bottom right figure shows the **shading predicted** by the binary model **with threshold 0.5**.

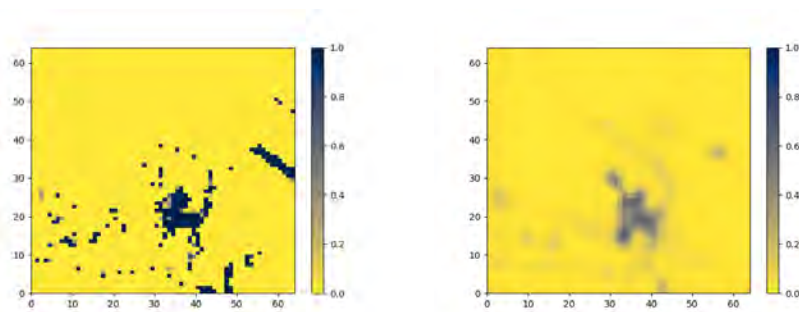
The model learning is basically achieved in the first 4 epochs. This is depicted in the left of Figure 1, and can be explained with the number of tiles used for training and the property of the U-Net to learn quickly with a small data set. One may observe in the right of Figure 1, the best results are obtained when the threshold is chosen at 0.35. If the choice is too low, the transitions can also be predicted as shaded. It is also remarkable that in this particular study the MSE-Loss performs best. In the literature[14][9], the cross-entropy loss is mostly used.

3.2 Shadow Depth Model

This model is more complex than the binary model, due to the nature of a multi-class net. Generally, it is expected that premature termination will not occur in this case, which can be also seen in Figure 3. Figure 4 shows that this net accurately calculates the shaded/non-shaded areas, but the transition between them is not sharp as in reality, which is why the result looks blurred. This can be remedied by an additional algorithm that sharpens the results.



■ **Figure 3** The figure shows the training metrics of the shadow depth model for 40 epochs. Unlike the binary model, the accuracy increases over the epochs so that the training was not terminated earlier.



■ **Figure 4** The left figure shows the original shadowing computed with the QGIS plugin. The right figure shows the shading predicted by the shadow depth model.

4 Conclusion

In this paper, the calculation of shading by a U-Net with residual layers was discussed and trained using selected test areas in Styria. As the results have shown, satisfactory values for the metrics, especially for accuracy, were obtained for both the binary net and the shadow depth net. This is an improvement over the current state of the art because there are currently no approaches that can predict the non-binary depth of the shadow. However, there are still a number of questions that are still open and are in need of further investigation. For example, the nets perform best with MSE loss as the training loss. However, the present state of affairs provides satisfactory results that may serve for further studies. Another aspect that needs to be further investigated and developed is the calculation time. Contrary to the literature, the approach shown is a factor of ten slower in the calculation of 10000 tiles than the QGIS plug-in with the traditional method.

In summary, with the approach of a U-Net as a basis for calculating the shadow depth, a suitable basis for further developments could be created.

References

- 1 Athanasios Angelis-Dimakis, Markus Biberacher, Javier Dominguez, Giulia Fiorese, Sabine Gadocha, Edgard Gnansounou, Giorgio Guariso, Avraam Kartalidis, Luis Panichelli, Irene Pinedo, et al. Methods and tools to evaluate the availability of renewable energy sources. *Renewable and sustainable energy reviews*, 15(2):1182–1200, 2011.
- 2 Sukriti Bhattacharya, Christian Braun, and Ulrich Leopold. An Efficient 2.5D Shadow Detection Algorithm for Urban Planning and Design Using a Tensor Based Approach. *ISPRS International Journal of Geo-Information*, 10(9):583, September 2021. doi:10.3390/ijgi10090583.
- 3 Andrew P Bradley. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern recognition*, 30(7):1145–1159, 1997.
- 4 Zoran Cuckovic. Enhancing terrain cartography with natural shadows, 2019. URL: <https://landscapearchaeology.org/2019/qgis-shadows/>.
- 5 Zoran Cuckovic. Terrain shading: a qgis plugin for modelling natural illumination over digital terrain models, 2021. URL: <https://github.com/zoran-cuckovic/QGIS-terrain-shading>.
- 6 Luciano da F. Costa. Further generalizations of the jaccard index. *CoRR*, abs/2110.09619, 2021. arXiv:2110.09619.
- 7 Lucas Fidon, Wenqi Li, Luis C Garcia-Peraza-Herrera, Jinendra Ekanayake, Neil Kitchen, Sébastien Ourselin, and Tom Vercauteren. Generalised wasserstein dice score for imbalanced multi-class segmentation using holistic convolutional networks. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: Third International Workshop, BrainLes 2017, Held in Conjunction with MICCAI 2017, Quebec City, QC, Canada, September 14, 2017, Revised Selected Papers 3*, pages 64–76. Springer, 2018.
- 8 Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. arXiv:1512.03385.
- 9 Fabian Isensee, Jens Petersen, Andre Klein, David Zimmerer, Paul F. Jaeger, Simon Kohl, Jakob Wasserthal, Gregor Koehler, Tobias Norajitra, Sebastian Wirkert, and Klaus H. Maier-Hein. nnu-net: Self-adapting framework for u-net-based medical image segmentation, 2018. arXiv:1809.10486.
- 10 Dominik Müller, Iñaki Soto-Rey, and Frank Kramer. Towards a guideline for evaluation metrics in medical image segmentation. *BMC Research Notes*, 15(1):1–8, 2022.
- 11 Jesús Polo, Nuria Martín-Chivelet, and Carlos Sanz-Saiz. BIPV Modeling with Artificial Neural Networks: Towards a BIPV Digital Twin. *Energies*, 15(11), 2022. doi:10.3390/en15114173.
- 12 P. Redweik, C. Catita, and M. Brito. Solar energy potential on roofs and facades in an urban landscape. *Solar Energy*, 97:332–341, 2013. doi:10.1016/j.solener.2013.08.036.
- 13 Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015. arXiv:1505.04597.
- 14 Nahian Siddique, Sidike Paheding, Colin P. Elkin, and Vijay Devabhaktuni. U-net and its variants for medical image segmentation: A review of theory and applications. *IEEE Access*, 9:82031–82057, 2021. doi:10.1109/ACCESS.2021.3086020.
- 15 Abdel Aziz Taha and Allan Hanbury. Metrics for evaluating 3d medical image segmentation: analysis, selection, and tool. *BMC medical imaging*, 15(1):1–28, 2015.
- 16 Sander van der Hoog. Deep learning in (and of) agent-based models: A prospectus, 2017. arXiv:1706.06302.
- 17 Yuan Yin, Vincent Le Guen, Jeremie Dona, Emmanuel de Bezenac, Ibrahim Ayed, Nicolas Thome, and Patrick Gallinari. Augmenting physical models with deep networks for complex dynamics forecasting. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12):124012, December 2021. doi:10.1088/1742-5468/ac3ae5.

Beware the Rise of Models When They Are Wrong: A Look at Heat Vulnerability Modeling Through the Lens of Sensitivity

Seda Şalap-Ayça¹ ✉ 🏠 

Department of Earth, Environmental, and Planetary Sciences, Brown University, Providence, RI, USA

Institute at Brown for Environment and Society, Brown University, Providence, RI, USA

Department of Earth, Geographic, and Climate Sciences, University of Massachusetts, Amherst, MA, USA

Erica Akemi Goto¹ ✉ 

Arizona Institute for Resilience, University of Arizona, Tucson, AZ, USA

Abstract

Extreme heat affects communities across the globe and is likely to increase as the climate changes; however, its consequences are not uniform. Geographically weighted regression is a useful modeling effort to understand the spatial linkage between various factors to heat-related casualty and supports decision-making in the spatial context. Still, as every complex spatial modeling approach, it is also bounded by uncertainty. Understanding model uncertainty and how this uncertainty is related to model input can be revealed by sensitivity analysis. In this study, we applied a spatial global sensitivity analysis to assess the model dynamics to address which input factors need to be prioritized in decision-making. A visual representation of the model's sensitivity and the spatial pattern of factor influence is an important step toward establishing a robust confidence mechanism for understanding heat vulnerability and supporting policy-making.

2012 ACM Subject Classification Information systems → Geographic information systems

Keywords and phrases heat vulnerability, uncertainty, sensitivity analysis

Digital Object Identifier 10.4230/LIPICs.GIScience.2023.64

Category Short Paper

Acknowledgements Seda Şalap-Ayça wants to thank Aykut Ayça for his discussion on selecting probability distribution functions.

1 Introduction and Background

Extreme heat causes injuries and fatalities in many regions of the U.S. Southwest region, especially during the summer months [12] [13]. Furthermore, extreme weather events like heat waves will likely increase with climate change [21]. However, studies have found that communities are not impacted the same by these extreme events [2] [9] as some communities are more vulnerable than others [9] [10].

There is a combination of factors that influence heat vulnerability, such as social (e.g., age and isolation[2][14][19]), economic (e.g., income and poverty[15]), health (e.g., pre-existing or chronic health conditions [15]), and environmental (e.g., lack of tree canopies or temperature [15]) factors. Scholars used these factors and applied various methods to measure heat vulnerability. Some of these methods were composite where the contributing factors are combined into an index [4][11][24] or regression which explains the relationship between

¹ corresponding author



independent and dependent variables [3],[16], or both [8]. Regression analysis draws a more reliable picture in terms of variables' influence on heat vulnerability compared to composite methods; however, the traditional regression methods do not address the spatial configuration of the factors, which can be solved by employing geographically weighted regression.

Despite the effort to address heat vulnerability, none of these methods are immune to uncertainty; as the presence and importance of uncertainty in spatial data is not new in data collection or GIS. Moreover, how the uncertainty is intertwined with vulnerability representation and its disproportionate impact on marginalized populations is an important dimension that is not addressed enough[6]. In the realm of policy decision-making models, when we contemplate the renowned aphorism of George Box, “all models are wrong but some are useful” in conjunction with Franklin’s [6] observation “poor data often disadvantages the disadvantaged” understanding uncertainty becomes more crucial in informed decisions and resource allocation. Therefore, in this study, we are interested in identifying how uncertainty in heat vulnerability related factors influences the prediction of health casualty. This research aims to advance the field of vulnerability to natural hazards and GIS by employing geographic weighted regression analysis coupled with sensitivity analysis.

2 Methodology

2.1 Data Acquisition and Variable Reduction

Our analysis area was the U.S. Southwest region, including Arizona, New Mexico, Texas, and Oklahoma states. The unit of analysis was chosen as the county level due to the availability of the data.

Our dataset was selected based on a thorough literature review of previous studies in the field and data availability. Our study combined social, economic, health, and environmental data as independent variables, county population (population 2015) and number of heat events as control variables, and casualty (fatalities and injuries) as the dependent variable. We used social and economic data from the 2015 American Community Survey 5-Year (ACS5), health data from both the ACS5 (2015) and the Global Health Exchange Data (GHDx) from 2014 and 2015, and environmental data from the National Oceanic and Atmospheric Administration (NOAA) from 2016 to 2020 and the Multi-Resolution Land Characteristics Consortium (MRLC) from 2016. Mortality and injury data were obtained from SHELDUS from 2016 to 2020. The initial dataset included 24 social, economic, health, and environmental variables.

We first tested the correlation between independent and dependent variables to reduce the number of independent variables. We then dropped independent variables whose relationship with the dependent variable was not statistically significant ($p - values > 0.05$). We also conducted a correlation matrix including all remaining independent variables and dropped one of the independent variables with a high correlation (> 0.7). Finally, we removed variables that had high spatial correlation. From our initial 24 independent variables, we ended up with 9 independent variables, which are elderly population, disabled population, black population, population with no car, unemployed population, number of months with a temperature higher than 38°C, and impervious surface and two control variables.

2.2 Geographically Weighted Regression for Heat Vulnerability

To understand the spatial relation between independent variables to dependent variables, we applied geographically weighted regression (GWR). GWR has been widely used over a decade to model the potential spatially varying relationships [1][5][23]. The model outcome

y_i (health-related casualty) can be expressed by

$$y_i = \beta_0(u_i, v_i) + \sum_k \beta_k(u_i, v_i)x_{ik} + \epsilon_i \quad (1)$$

where y_i represents the dependent variable, k are independent variables, β is the parameter to be estimated, ϵ is the error term. (u_i, v_i) denotes the coordinates of the i^{th} feature and $k(u_i, v_i)$ is a realization of the continuous function $k(u, v)$ at feature i .

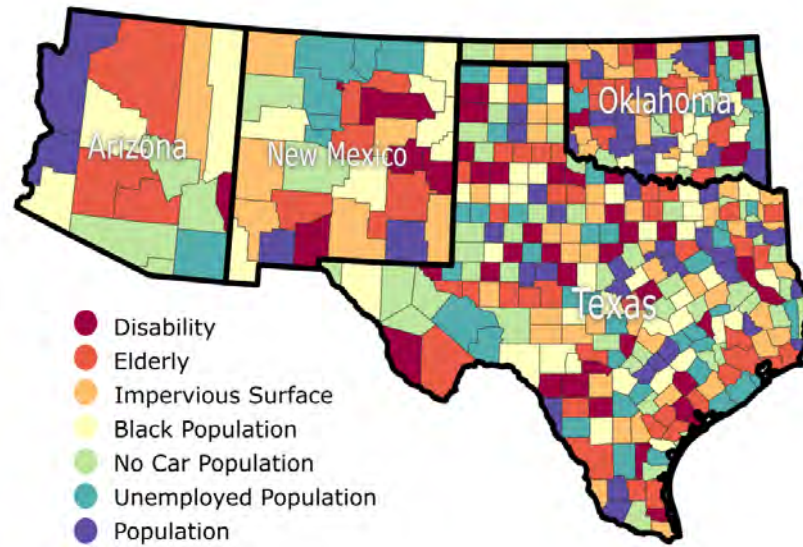
2.3 Global Sensitivity Analysis for Heat Vulnerability GWR

Global Sensitivity analysis (GSA) is a forward looking approach to modeling to understand the linear (individual) and nonlinear (interaction) relationship between input variables and the output of the model [18]. In this study, since we are focusing on GWR's sensitivity, our focus is on which independent variable(s) influence most the prediction of causality. GSA starts with generating random samples which are used to replicate the behavior of the input set when the model is run multiple times. These sample sets mimic the original probability distribution function (pdf) of the input variables. Therefore, we conducted a systematic analysis to acquire the pdf of each variable before generating the samples. Due to limited information about some factors' priori distribution, GSA is only conducted for 7 variables (disability, elderly, impervious surface, black population, population with no car, unemployed, 2015 county population), whereas all 9 are included in the GWR. Once the pdfs are determined, the following framework is applied:

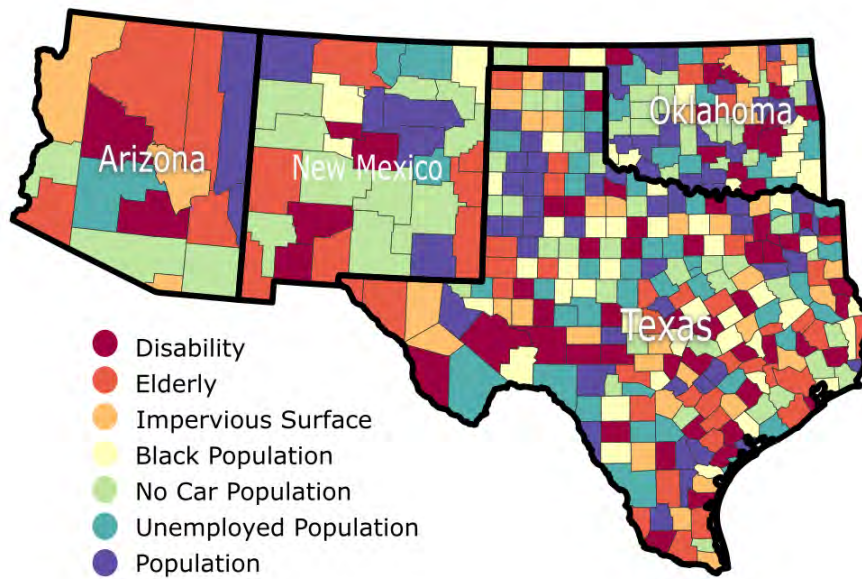
1. For each parameter, we generated 2048 random variables. This number is based on the experimental example set N (2^7) and number of model inputs $D(7)$, which yields $N(2D + 2) = 2048$ samples.
2. Prepare the sample set for GWR input using python based pandas library [22]
3. Run GWR model 2048 times on county scale with randomly generated sample set of independent variables
4. Exporting GWR output for GSA
5. Implementing GSA using SALib package in python [20] [7]

3 Results and Discussion

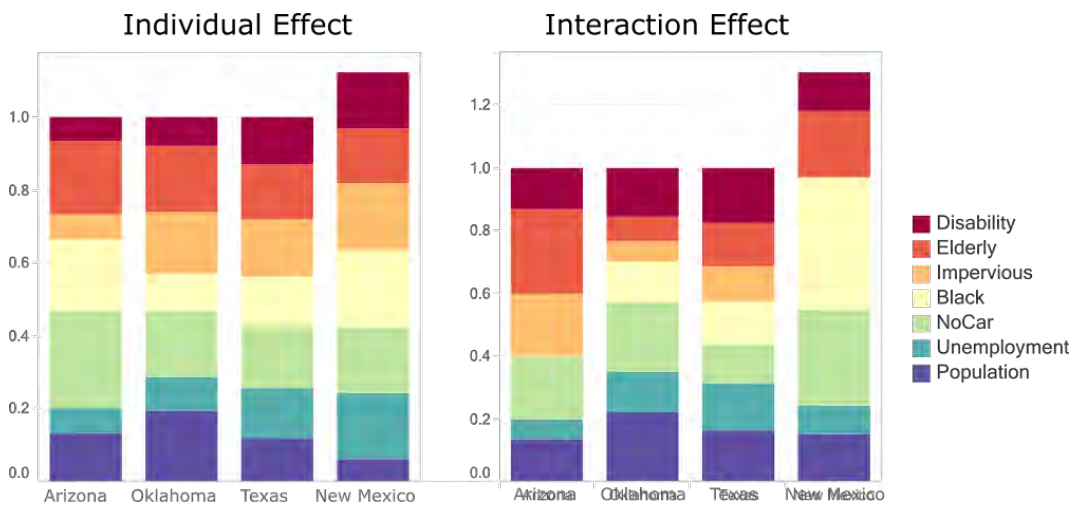
The result of the GSA has been visualized as the most influential variable for the individual (Figure 1) and interaction effects' (Figure 2) influences in terms of the model's explanatory power for each county. These maps represent variables contributing the most to the model's variability or uncertainty. For each input variable that is fed into GSA, the analysis produces a unique GSA map. In order to reduce the visual load of the GSA output (7 individual and 7 interaction effects map), self-organizing map-based exploratory analysis [17] has been used where the neural networks evaluate the similarities among indices per future and results in clusters where patterns are dominant. While Figure 1 shows us where individual variances of each variable affect the heat-related casualty uncertainty, Figure 2 depicts the interaction effect influence on the model uncertainty. For example, when we look at Arizona State, GWR output is most sensitive to any small variation in no car variable (observed in 3 counties) when each independent variable is singly treated (Figure 3). However, as the spatial complex nature of these variables plays an important role in GWR prediction, we can see an increase in disability and elderly variables when the interaction among parameters is considered. This means individual effects will not be enough to see the whole picture when we try to understand model dependencies. Also, as we can see, the influential variables vary among the four states. When heat vulnerability modeling efforts are in action, each state might prioritize its resources depending on how these variables are distributed.



■ **Figure 1** Most influential factors to the model uncertainty based on Individual Effects of Input Variables to Predicted Heat-related Casualty.



■ **Figure 2** Most influential factors to the model uncertainty based on Interaction Effects of Input Variables to Predicted Heat-related Casualty.



■ **Figure 3** Frequency distribution of individual effect dominant factors per state.

4 Conclusion and Future Work

The geographic focus is crucial for equitable risk planning, resilience strategies, and response to heat risk. Moreover, it can be used to communicate results and support decision-makers. Data acquisition is the most time and effort-consuming part of the spatial decision-making process; but crucial as the interaction of variables produces different results. Considering the unavoidable uncertainty, it is important to know the models' weaknesses and strengths and the spatial variability of the results so that the resource allocation can be optimum. Moreover, heat vulnerabilities indicated by dominant factors depicted in Figures 1 and 2, can help decision makers and modelers to prioritizing resources. This effort will help us identify the influential variables and where they cluster as an initial step and can be followed by the involvement and insight of the communities which need to be a part of the solution.

References

- 1 Chris Brunsdon, Stewart Fotheringham, and Martin Charlton. Geographically weighted regression-modelling spatial non-stationarity. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 47(3):431–443, 1998. URL: <http://www.jstor.org/stable/2988625>.
- 2 Susan L Cutter, Bryan J Boruff, and WL Shirly. Social science quarterly. *Soc. Vulnerability Environ. Hazards*, 84:242–261, 2003.
- 3 Yaella Depietri, Torsten Welle, and Fabrice G. Renaud. Social vulnerability assessment of the cologne urban area (germany) to heat waves: links to ecosystem services. *International Journal of Disaster Risk Reduction*, 6:98–117, 2013.
- 4 Abbas El-Zein and Fahim N Tonmoy. Assessment of vulnerability to climate change using a multi-criteria outranking approach with application to heat stress in sydney. *Ecological Indicators*, 48:207–2017, 2015.
- 5 A. Fotheringham, Chris Brunsdon, and Martin Charlton. *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*. Wiley, 2002.
- 6 Rachel Franklin. Quantitative methods i: Reckoning with uncertainty. *Progress in Human Geography*, 46(2):689–697, 2022.
- 7 Jon Herman and Will Usher. SALib: An open-source python library for sensitivity analysis. *The Journal of Open Source Software*, 2(9), January 2017.

- 8 Daniel P. Johnson, Austin Stanforth, Vijay Lulla, and George Luber. Developing an applied extreme heat vulnerability index utilizing socioeconomic and environmental data. *Applied Geography*, 35(1):23–31, 2012.
- 9 Shirley Laska and Betty Hearn Morrow. Social vulnerabilities and hurricane katrina: an unnatural disaster in new orleans. *Marine technology society journal*, 40(4), 2006.
- 10 K. Lieberknecht, D. Zoll, and K. Castles. Hurricane harvey: equal opportunity storm or disparate disaster? *Local Environment*, 2(26), 2021.
- 11 Francisco de la Barrera Luis Inostroza, Massimo Palme. A heat vulnerability index: Spatial patterns of exposure, sensitivity and adaptive capacity for santiago de chile. *PLOS ONE*, 2016.
- 12 Alex Nguyen and Erin Douglas. Texas heat-related deaths reached a two-decade high in 2022 amid extreme temperatures. *CHRON*, 2022.
- 13 AZHS Arizona Department of Health Services. Extreme weather & public health. URL: <https://www.azdhs.gov/preparedness/epidemiology-disease-control/extreme-weather/heat-safety/index.php#heat-home>.
- 14 D. Reckien. What is in an index? construction method, data metric, and weighting scheme determine the outcome of composite social vulnerability indices in new york city. *Regional Environmental Change*, 18:1439–1451, 2018.
- 15 C. Rinner, D. Patychuk, K. Bassil, S. Nasr, S. Gower, and M. Campbell. The role of maps in neighborhood-level heat vulnerability assessment for the city of toronto. *Cartography and Geographic Information Science*, 1(37), 2010.
- 16 Samain Sabrin, Maryam Karimi, Md Golam Rabbani Fahad, and Rouzbeh Nazari. Quantifying environmental and social vulnerability: Role of urban heat island and air quality, a case study of camden, nj. *Urban Climate*, 34, 2020.
- 17 Seda Şalap-Ayça. Self-organizing maps as a dimension reduction approach for spatial global sensitivity analysis visualization. *Transactions in GIS*, 26(4):1718–1734, 2022.
- 18 Seda Şalap-Ayça, Piotr Jankowski, Keith C Clarke, Phaedon C Kyriakidis, and Atsushi Nara. A meta-modeling approach for spatio-temporal uncertainty and sensitivity analysis: an application for a cellular automata-based urban growth and land-use change model. *International Journal of Geographical Information Science*, 32(4):637–662, 2018.
- 19 S. Sheridan T. Dolney. The relationship between extreme heat and ambulance response calls for the city of toronto, ontario, canada. *Environmental Research*, 100:94–103, 2006.
- 20 Iwanaga Takuya, William Usher, and Jonathan Herman. Toward SALib 2.0: Advancing the accessibility and interpretability of global sensitivity analyses. *Socio-Environmental Systems Modelling*, 4:18155, May 2022. doi:10.18174/sesmo.18155.
- 21 Michael Wehner, Sonia Seneviratne, Xuebin Zhang, Muhammad Adnan, Wafae Badi, Claudine Dereczynski, Alejandro Di Luca, Subimal Ghosh, Iskhaq Iskandar, James Kossin, et al. Weather and climate extreme events in a changing climate. In *AGU Fall Meeting Abstracts*, volume 2021, pages U13B–11, 2021.
- 22 Wes McKinney. Data Structures for Statistical Computing in Python. In Stéfan van der Walt and Jarrod Millman, editors, *Proceedings of the 9th Python in Science Conference*, pages 56–61, 2010.
- 23 David C. Wheeler and Antonio Páez. Geographically weighted regression. In Manfred M. Fischer and Arthur Getis, editors, *Handbook of Applied Spatial Analysis: Software Tools, Methods and Applications*, pages 461–486. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010.
- 24 Tanja Wolf and Glenn McGregor. The development of a heat wave vulnerability index for london, united kingdom. *Weather and Climate Extremes*, 1:59–68, 2013.

From Change Detection to Change Analytics: Decomposing Multi-Temporal Pixel Evolution Vectors

Victoria Scherelis   

Zurich University of Applied Sciences, Wädenswil, Switzerland
University of Zurich, Switzerland

Patrick Laube  

Zurich University of Applied Sciences, Wädenswil, Switzerland
University of Zurich, Switzerland

Michael Doering  

Zurich University of Applied Sciences, Wädenswil, Switzerland

Abstract

Change detection is a well-established process of detecting spatial and temporal changes of entities between two or more timesteps. Current advancements in digital map processing offer vast new sources of multitemporal geodata. As the temporal aspect gains complexity, the dismantling of detected changes on a pixel-based scale becomes a costly undertaking. In efforts to establish and preserve the evolution of detected changes in long time series, this paper presents a method that allows the decomposition of pixel evolution vectors into three dimensions of change, described as directed change, change variability, and change magnitude. The three dimensions of change compile to complex change analytics per individual pixels and offer a multi-faceted analysis of landscape changes on an ordinal scale. Finally, the integration of class confidence from learned uncertainty estimates illustrates the avenue to include uncertainty into the here presented change analytics, and the three dimensions of change are visualized in complex change maps.

2012 ACM Subject Classification Information systems → Spatial-temporal systems

Keywords and phrases Digital map processing, spatio-temporal modelling, land-use change

Digital Object Identifier 10.4230/LIPIcs.GIScience.2023.65

Category Short Paper

Funding Swiss National Science Foundation under Grant number 200021_188692/1.

Acknowledgements We thank Dominic Lüönd for his support with Figure 3.

1 Introduction

Change detection (CD) is the process of capturing the spatial and temporal changes of individual pixels, objects, or larger phenomena. Requiring a minimum of two timesteps, the most common types of change detection include pixel-based (PBCD) and object-based change detection (OBCD). These differ in that PBCD is focused on pixel-wise changes and usually on the spectral value of the individual pixel with no spatial relevance to its neighbors [7], and OBCD on the object represented by the pixels and grouping/segmenting the pixels into clusters of their respective categories [1]. Within the very mature field of CD, many reviews have been published to organize the different CD types [1, 7] and the vast methods and techniques used to detect changes within the various categories [2]. Current advances in CD are mostly built on existing foundations and are focused on the development of automatic change detection algorithms tailored to specific topics and based on complex neural networks or other deep learning algorithms [5].



© Victoria Scherelis, Patrick Laube, and Michael Doering;
licensed under Creative Commons License CC-BY 4.0

12th International Conference on Geographic Information Science (GIScience 2023).

Editors: Roger Beecham, Jed A. Long, Dianna Smith, Qunshan Zhao, and Sarah Wise; Article No. 65; pp. 65:1–65:6



Leibniz International Proceedings in Informatics

Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

A large application field of CD is in detecting land use and land cover (LULC) changes [4, 13], including habitat changes in riverine environments [11], the topic of the here presented study. Both PBCD and OBCD are used in LULC, mostly identifying changes between only two or three timesteps and focused on satellite data or aerial imagery [4, 13]. While some studies exist that incorporate multiple timesteps, such as in Tonolla et al. [11], there is a general lack in CD application with a large temporal depth of more than three timesteps, specifically in applying CD methods between individual timesteps of such large multitemporal data. In addition, the majority of CD methods are focused on remote sensing based data which reflect changes only since the 1940s and may already represent disrupted environments. Historical maps offer a unique perspective on pre-digital and perhaps pre-modified times, yet the application of CD methods on historical map sources is poorly represented within the literature with only few scattered examples [8]. The establishment and preservation of CD in longer time series is a costly undertaking and hence methods analytically exploiting such data are rarely seen. However, efforts in conceptualizing PBCD are well underway as space-time relationships are at the core of GIS research [3]. With rapid advancements in utilizing machine learning based digital map processing [9], more extensive sources of time series become available, offering new opportunities for exploiting such data, as shown in this study.

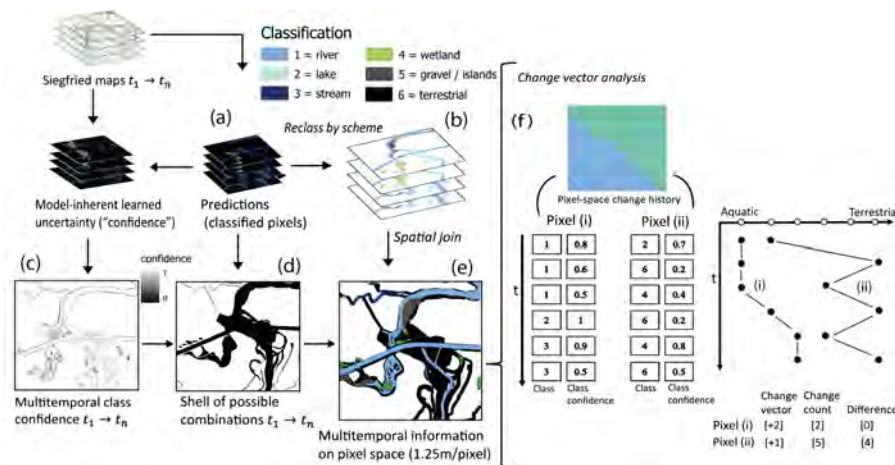
Visualization of the CD between two or three timesteps are shown in several studies in forms of change maps. These change maps of very few timesteps visualize changes on a binary scale of either ‘change’ or ‘no change’ [5] or show separate maps of the conditions per timestep [4, 13]. Visualization of larger temporal extent tend to show the change event of largest magnitude of a pixel [6], resulting in a rather static perspective of the observed changes. The visualization approaches of CD within the literature poorly represent the potentially very rich dynamic nature of the changes through time.

This paper presents a novel approach for the quantitative analysis of rich raster data time series, allowing an in depth analysis of detailed change evolution vectors per individual pixel. The experimental part of the paper illustrates the methods for multitemporal PBCD from Switzerland’s historical maps. The pixel-wise change analytics incorporate 6 timesteps between 1876-1946. Change is decomposed into three dimensions that are visualized in three separate maps. A pixel flow chart illustrates changes between landuse classes. In addition, the change analytics also incorporate class confidence (learned uncertainty) per pixel, as model-inherent uncertainty is introduced in the extraction process from historical maps.

2 Methods

Fig. 1 illustrates the conceptual workflow from the historical map inputs to the pixel-wise change analytics. The components of the workflow are described in detail below.

Data pre-processing. The data inputs used in this study are classified pixel clusters extracted from the historical Siegfried map series of Switzerland. Based on training data, hydrological features (i.e. rivers, wetlands) are defined and grouped by certain criteria of their appearance on the maps [10] for extraction. The pixels representing these features are then extracted from the maps by deep learning algorithms [12] which output predictions in form of classified pixel clusters (Fig. 1a). A type of model-inherent uncertainty based on learned confidence estimates (LCE) are an additional output from the extraction process (see [12] for details) and the basis to determine class confidence/ uncertainty in this study. For further application in terms of habitat changes in ecohydrological environments, the predictions are reclassified based on a hierarchical classification scheme from aquatic to



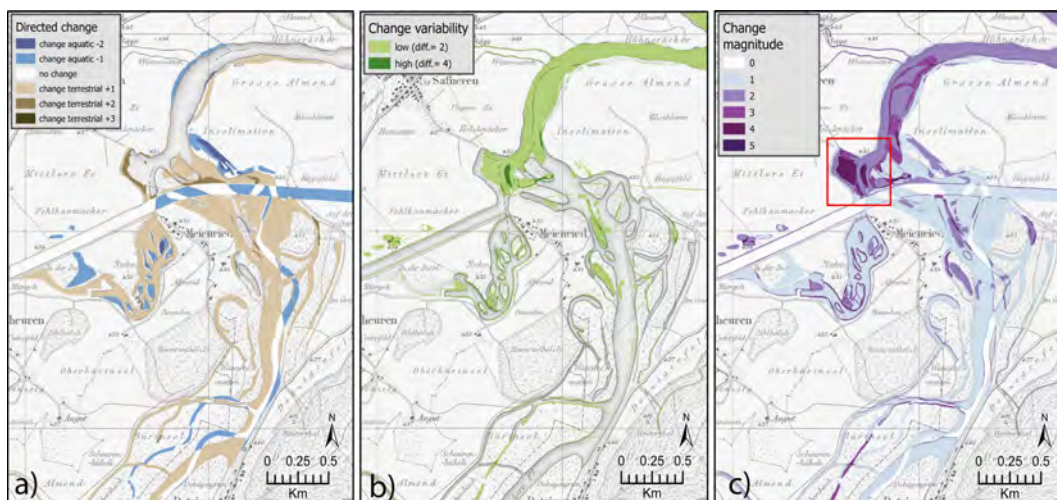
■ **Figure 1** Conceptual workflow for multi-temporal change detection and derivation of pixel-wise change analytics ($t = \text{time}$).

terrestrial classes to determine their directed transition between the class types (Fig. 1b). The terrestrial class 6 presents any pixel unclassified from the predictions. Although the methods are presented here in the context of habitat succession, they are applicable for any geodata time series with attributes on an ordinal scale.

Integration of model-inherent uncertainty. Learned confidence estimates (LCE) show class uncertainty per pixel of the four hydrological feature classes of rivers, wetlands, lakes, and streams. Each pixel thus has 4 confidence values between 0-1 which depicts the models' uncertainty of that pixels predicted class. To integrate the LCE, the classified pixels in the predictions and the uncertainty estimates of the class identified in the associated prediction, hereafter class confidence, are extracted by a conditional evaluation. For each pixel in each timestep, the associated class confidence is then represented in the change analytics. Based on the available timesteps, the mean of the class confidence per pixel is calculated to depict the average multitemporal class confidence per pixel from $t_1 \rightarrow t_n$ (Fig. 1c). Note, LCE were available only for the first four timesteps. Thus, the represented class confidence of the time series are based on the average of the first four timesteps.

Integration of multitemporal information. For computational purposes and to avoid large amounts of terrestrial pixels, a shell of all possible multitemporal combinations was derived. The shell can be described as a sparsely populated array which occupies a pixel space when one classified pixel within any of the timesteps from $t_1 \rightarrow t_n$ occupies that pixel space, either from the predictions or from the LCE (Fig. 1d). Based on their spatial distribution, the information of the reclassified pixels were joined on to the shell. From a multitemporal perspective, not all pixel spaces have an associated class within all timesteps as the hydrological features represented by the classes change through time. Thus, pixel-spaces with no associated class in a given timestep are assigned the terrestrial class. The new dataset then holds information on the class of each pixel-space through time (Fig. 1e).

Change analytics and change vector analysis. A common CD method is the change vector analysis which can be described as the difference between the spectral pixel vector of two images [13]. The method is adapted and modified to identify the change direction and magnitude of class memberships per pixel in a multitemporal dataset with 6 timesteps (Fig. 1f). On an ordinal scale between aquatic and terrestrial, the *directed change* between two timesteps is evaluated as $DC = \sum_{n=1}^{i=1} \frac{x_{i+1} - x_i}{|x_{i+1} - x_i|}$, where DC is the change vector and x



■ **Figure 2** Visualization of the three dimensions of change in the pixel-wise change analytics, applied on a region of the Aare river in Switzerland. The background shows the 1876 Siegfried map before channelization. The red outline in (c) highlights a section for which the pixel flow paths through time are shown in Fig.3. Data source swisstopo.ch.

is a pixel of timestep i . In this equation, the difference between x_{i+1} and x_i is divided by the absolute value of that difference. This results in a value of +1 if x_{i+1} is greater than x_i (i.e., a change towards terrestrial) and -1 if x_{i+1} is less than x_i (i.e., a change towards aquatic). The expression evaluates to 0 if there is no change. The sum of the expression from $t_1 \rightarrow t_n$ then results in a change vector presenting multitemporal *directed changes*, the first dimension of the here presented change analytics.

The second dimension includes an evaluation of *change magnitude* (CM) of all changes which occurred between consecutive timesteps, described by $CM = \sum_{n=1}^{i=1} |x_{i+1} - x_i| > 0$. For each timestep i , the expression $|x_{i+1} - x_i| > 0$ evaluates to TRUE if there was a class change in x between i and $i + 1$, and FALSE if otherwise. The absolute values ensure a positive expression regardless of the direction of change.

Lastly, true changes with direction are differentiated from *change variability*, where frequent changes between individual classes occur with no clear direction. This difference is evaluated by $CVar = [CM] - [|DC|]$ and depicts the third dimension of change. As exemplified in pixel (i) and (ii) in Figure 1(f), large differences between CM and $|DC|$ indicate that the class membership of the particular pixel frequently fluctuated between specific classes, whereas small to no differences indicate true directional change of the class towards aquatic or terrestrial. The change analytics per pixel can then be visualized by the calculated difference to show pixels with multitemporal variability. Where $CVar > 0$, the DC values are visualized to show the relative magnitude and direction of pixels that observed multitemporal change.

3 Results and Discussion

The detailed change analytics per individual pixel showed regions of directional change as well as regions of variability over the investigated time series of 1876 to 1946, with a timestep roughly every 14 years. The presented methods allowed the decomposition of pixel evolution vectors into three dimensions of change. Fig. 2 visualizes these dimensions of change for a

■ **Table 1** Areas of directed change and change variability with averaged class confidence.

(1.25m/pixel)	Area (m^2) full map sheet	Average confidence
High variability	94'139	0.65
Low variability	2'700'396	0.84
Change aquatic	881'232	0.93
Change terrestrial	2'294'343	0.91

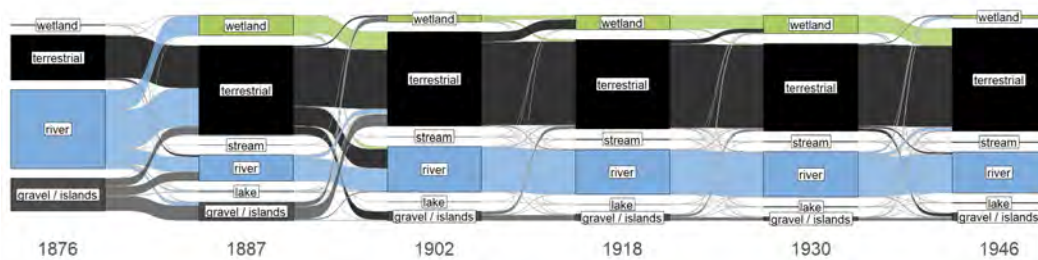
section of the map sheet under study. Fig. 2(a) depicts the dimension of directed change towards terrestrial or aquatic classes with the respective magnitude, regions where no changes occurred are shown in white. Fig. 2(b) illustrates change variability, the change dimension where pixels observed frequent alternations between specific classes (high). Some (low) change variability is observed when the classes alternate at least twice over the time series, differences of three ($\text{diff}=3$) or change variability higher than four variations ($\text{diff}>4$) were not observed. Fig. 2(c) visualizes the change magnitude observed per pixel throughout the time series, the third dimension of change.

Table 1 summarizes the areas which observed directed change and change variability, with the averaged multitemporal class confidence per pixel from the change analytics, for the map sheet TA 124 under study. The results show that the majority of pixels changed towards the terrestrial class or observed low variability, meaning that class alternations were only observed once or twice in those pixels. The class confidence was relatively low for regions that observed high variability, indicating that high variability in class changes also introduces larger uncertainties in class confidence as the class membership defined per timestep are less certain and could likely be identified as other classes in the specific timestep.

The change analytics enabled us to quantify the path of changes of each individual pixel. With six class types and 6 timesteps, over 1000 combinations of class changes were observed within the change analytics. Figure 3 illustrates the observed paths of individual pixels and their classes through time. The region of path combinations shown is outlined in red in Fig. 2c. Overall, the change history shows that terrestrial and lake classes steadily increased over time. Rivers and streams overall decreased with small increases between individual timesteps while gravel deposits and islands steadily decreased. Wetlands varied throughout the time series but generally increased over time.

4 Conclusion and Outlook

In this article, we proposed a novel approach to investigate changes in raster based data time series, generating multi-dimensional change analytics per individual pixels. The here presented methods allowed the decomposition of pixel evolution vectors into three dimensions



■ **Figure 3** Flow paths of pixels and their observed classes through time (R package parcats).

of change: directed change, change variability, and change magnitude. The change analytics can be visualized by complex change maps depicting the three dimensions of change observed per pixel-space. With a unique application of PBCD on historical map sources, the change analytics included 6 timesteps and incorporated class confidence per pixel. Overall, the here presented methods offer differentiated insights in complex change dynamics, and do so considering uncertainty.

In terms of future work, we aim to incorporate the LCE for the remaining timesteps and, instead of viewing the averaged multitemporal class confidence, integrate the class confidence per class type and timestep into the pixel level change analytics. In addition, to make further use of the detail gained by the change histories, we intend to incorporate overall relative change for *DC* and overall absolute change for *CM* to capture the absolute magnitude of overall observed changes. Further timesteps and other map sheets will be investigated to test the robustness of the CD approach. In general, the change analytics described in this paper and its visualization of complex spatio-temporal data has great application potential to other fields detecting changes in multitemporal time series.

References

- 1 P Aplin and GM Smith. Advances in object-based image classification. *The Int. Archives of the Photogrammetry, Remote Sensing and Spatial Info. Sciences*, 37(B7):725–728, 2008.
- 2 Priti Attri, Smita Chaudhry, and Subrat Sharma. Remote Sensing & GIS based Approaches for LULC Change Detection – A Review. *Remote Sensing*, 2015.
- 3 A. Comber and M. Wolter. Considering spatiotemporal processes in big data analysis: Insights from remote sensing of land cover and land use. *Transactions in GIS*, 23(5):879–891, 2019.
- 4 Limin Dai, Shanlin Li, Bernard J. Lewis, Jian Wu, Dapao Yu, Wangming Zhou, Li Zhou, and Shengnan Wu. The influence of land use change on the spatial–temporal variability of habitat quality between 1990 and 2010 in Northeast China. *J. of Forestry Res.*, 30(6):2227–2236, 2019.
- 5 Tianqi Gao, Hao Li, Maoguo Gong, Mingyang Zhang, and Wenyuan Qiao. Superpixel-based multiobjective change detection based on self-adaptive neighborhood-based binary differential evolution. *Expert Systems with Applications*, 212:118811, 2023.
- 6 T. Hermosilla, M. A. Wolter, J. C. White, N. C. Coops, G. W. Hobart, and L. B. Campbell. Mass data processing of time series Landsat imagery: pixels to data products for forest monitoring. *Int. J. of Digital Earth*, 9(11):1035–1054, 2016.
- 7 Masroor Hussain, Dongmei Chen, Angela Cheng, Hui Wei, and David Stanley. Change detection from remotely sensed images: From pixel-based to object-based approaches. *ISPRS J. of Photogrammetry and Remote Sensing*, 80:91–106, 2013.
- 8 L. A. James, M. E. Hodgson, S. Ghoshal, and M. M. Latiolais. Geomorphic change detection using historic maps and DEM differencing: The temporal dimension of geospatial analysis. *Geomorphology*, 137(1):181–198, 2012.
- 9 Chenjing Jiao, Magnus Heitzler, and Lorenz Hurni. A survey of road feature extraction methods from raster maps. *Transactions in GIS*, 25(6):2734–2763, 2021.
- 10 Victoria Scherelis, Michael Doering, Marta Antonelli, and Patrick Laube. Hydromorphological Information in Historical Maps of Switzerland: From Map Feature Definition to Ecological Metric Derivation. *Annals of the Am. Asso. Geographers*, pages 1–18, 2023.
- 11 Diego Tonolla, Martin Geilhausen, and Michael Doering. Seven decades of hydrogeomorphological changes in a near-natural and a hydropower-regulated pre-Alpine river floodplain in Western Switzerland. *Earth Surface Proc. and Landforms*, page 5017, 2020.
- 12 Sidi Wu, Magnus Heitzler, and Lorenz Hurni. Leveraging uncertainty estimation and spatial pyramid pooling for extracting hydrological features from scanned historical topographic maps. *GIScience & Remote Sensing*, 59(1):200–214, 2022.
- 13 Song Xiaolu and Cheng Bo. Change detection using change vector analysis from landsat tm images in wuhan. *Procedia Environmental Sciences*, 11:238–244, 2011.

How to Count Travelers Without Tracking Them Between Locations

Nadia Shafaeipour ✉ 

Faculty of Geo-Information Science and Earth Observation (ITC), University of Twente, Enschede, The Netherlands

Maarten van Steen ✉ 

Digital Society Institute (DSI), University of Twente, Enschede, The Netherlands

Frank O. Ostermann ✉ 

Faculty of Geo-Information Science and Earth Observation (ITC), University of Twente, Enschede, The Netherlands

Abstract

Understanding the movements of travelers is essential for sustainable city planning, and unique identifiers from wireless network access points or smart card check-ins provide the necessary information to count and track individuals as they move between locations. Nevertheless, it is challenging to deal with such uniquely identifying data in a way that does not violate the privacy of individuals. Even though several protection techniques have been proposed, the data they produce can often still be used to track down specific individuals when combined with other external information. To address this issue, we use a novel method based on encrypted Bloom filters. These probabilistic data structures are used to represent sets while preserving privacy under strong cryptographic guarantees. In our setup, encrypted Bloom filters offer statistical counts of travelers as the only accessible information. However, the probabilistic nature of Bloom filters may lead to undercounting or overcounting of travelers, affecting accuracy. We explain our privacy-preserving method and examine the accuracy of counting the number of travelers as they move between locations. To accomplish this, we used a simulated subway dataset. The results indicate that it is possible to achieve highly accurate counting while ensuring that data cannot be used to trace and identify an individual.

2012 ACM Subject Classification Security and privacy → Domain-specific security and privacy architectures

Keywords and phrases Privacy preservation, encrypted Bloom filters, traveler counting, subway networks

Digital Object Identifier 10.4230/LIPIcs.GIScience.2023.66

Category Short Paper

Supplementary Material *Software (Source Code)*: <https://github.com/Nadia-Shafaeipour/Counting-travelers-BFs>, archived at [swh:1:dir:58544f9167cea9d5e0f8b973178a59bbca8768aa](https://www.swh.io/dir/58544f9167cea9d5e0f8b973178a59bbca8768aa)

Funding *Nadia Shafaeipour*: This work was supported by Dutch Research Council(NWO).

1 Introduction

As urbanization continues to rise, the usage of public transportation modes is increasing. To implement policies that increase the sustainability of urban transportation systems, a deeper understanding of travel patterns is essential. Traditional surveys and travel diaries require significant effort and provide only snapshots [2]. Various more recent technologies allow automated counting, e.g., Bluetooth and Wi-Fi detection systems and automated fare collection systems. Information gathered by these systems has proved to be helpful in improving security, physical activity, traffic safety, public transportation, communication infrastructure [7, 5], and the overall quality of life for citizens.



© Nadia Shafaeipour, Maarten van Steen, and Frank O. Ostermann; licensed under Creative Commons License CC-BY 4.0

12th International Conference on Geographic Information Science (GIScience 2023).

Editors: Roger Beecham, Jed A. Long, Dianna Smith, Qunshan Zhao, and Sarah Wise; Article No. 66; pp. 66:1–66:6

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

However, counting travelers at specific locations by using smart-card IDs allows tracking their movements between locations. It is, therefore, a sensitive issue, especially if it allows monitoring travelers over an extended period of time: the trade-off for valuable insights into movement patterns is an infringement upon their privacy. It has been shown that a few points are enough to identify individual travelers with simple anonymization and persistent identifiers [4].

To prevent such situations, various regulations have been adopted, including Europe’s General Data Protection Regulation (GDPR) [6], which requires parties to obtain explicit consent before collecting and using personal information. Obtaining explicit consent may reduce the completeness of the data collection, which can introduce bias and reduce the representativeness of the results. Even if consent is granted, individuals must trust that their data will be used responsibly and not misused for other purposes.

For these reasons, we challenge the feasibility of robust privacy protection within a system that relies on identifying travelers to count them. Instead, we propose an alternative system that offers statistical counts of travelers as the only accessible information. To implement such a system, we propose to use Bloom filters, which are probabilistic data structures that support set operations, in combination with homomorphic encryption, which is a type of encryption that allows performing operations on encrypted data. We envision a system that provides reliable counts of travelers moving between locations as the only retrievable information [10].

In this paper, we explain and briefly evaluate our privacy-preserving method for its accuracy in counting travelers moving between locations, with the aim to show its principal working. As a case study, we consider a subway network where travelers utilize smart-card technology to check in and out of the transportation system. To accomplish this, we use a synthetic dataset that is accurate in representing the characteristics of a real-world subway dataset. The results demonstrate the effective combination of Bloom filters and homomorphic encryption in accurately counting travelers between locations while preserving individual privacy. This finding paves the way for expanding the analysis to include multiple locations within the subway network. Our research carries significant implications for enhancing public transportation efficiency and safeguarding user privacy.

2 System model

Our example proof-of-concept assumes a subway network with an automatic fare collection system. Subway networks usually consist of lines that connect specific origin and destination stations. For each station A , we assume there is a set of n_A scanners $\mathcal{S}_A = \{s_1^A, \dots, s_{n_A}^A\}$, which are used by travelers to check in and out. In our model, we trust the sensors, but not the centralized server. For this reason, we first let a sensor collect detections to then send this collection in encrypted form to the server. The time during which detections are collected and aggregated before sending them to the server is called an **epoch**. Typically, an epoch lasts 5 minutes. As we will discuss in detail below, the server can operate on the encrypted collections of detections, but cannot reconstruct individual detections themselves.

A scanner $s \in \mathcal{S}_A$ reads a card’s unique identifier cid . Each card reading belongs to an epoch $e \in \mathcal{E}$ corresponding to its timestamp t , such that $t_{start}(e) \leq t < t_{end}(e)$, where t_{start} and t_{end} mark the beginning and the end of an epoch and \mathcal{E} denotes the set of all such epochs. A detection is thus a triplet (cid, s, e) , representing a card uniquely identified by its identifier cid , read by scanner s during epoch e . By $\mathcal{D}_{s,e}$, we denote the set containing all the identifiers detected by a scanner s during an epoch e . Let \mathcal{D}_e^A denote the set of all identifiers detected by *any* scanner at A during epoch e : $\mathcal{D}_e^A = \cup_{s \in \mathcal{S}_A} \mathcal{D}_{s,e}$.

Using collections of detections provides a powerful mechanism for counting travelers. One simple example is that the size of a set $\mathcal{D}_{s,e}$ indicates the number of travelers who passed the scanner s during the epoch e . More interestingly, for two different stations, the size of the set $\mathcal{D}_{e_1}^A \cap \mathcal{D}_{e_2}^B$ represents the number of travelers who were first detected at A during epoch e_1 and subsequently detected at B during epoch e_2 , where e_2 occurs after e_1 .

3 Method

3.1 Bloom filter

The problem with using sets of detections is that they still contain the card identifiers for anyone to see who has access to those sets. This issue can be addressed by using a *representation* for sets, called **Bloom filters** [3]. A Bloom filter has the property that it allows only for membership tests. In other words, the only way to discover which card identifier is stored, is to go over the entire list of possible card identifiers and check for each one of them which identifier the membership tests succeed. Although this already poses a potentially tremendous computational burden for discovering detected identifiers, it is not enough to prevent finding identifiers. To understand how encryption, combined with Bloom filters, can prevent such a discovery, we must first explain what they are.

A Bloom filter is implemented as a binary vector of m bits, initially all set to zero. Adding an element to the set involves hashing it with k different hash functions, each returning a position in the vector. Those bits are then set to 1. To determine whether an element is in the set, the same hash functions are applied, and the corresponding bits in the vector are checked. When each bit is also 1, the element is considered to be in the set. An important observation is that there is a chance that two different elements will see exactly the same bits being set to 1. As a consequence, a membership test may return a *false positive*: the element for which the test is computed is factually *not* in the set represented by the Bloom filter. It is for this reason that Bloom filters are said to be probabilistic data structures. Given the maximum acceptable probability p for false positives, along with the desired number n of elements to be stored, one can compute the minimal length m of a Bloom filter, as well as the minimal number k of hash functions to use: $m = -\frac{n \cdot \ln p}{(\ln 2)^2}$ and $k = \frac{m}{n} \cdot \ln 2$.

The size of the set represented by a Bloom filter (i.e., its *cardinality* c) can be estimated when knowing only k , m , and the number t of bits that are set to 1 [8]:

$$c = -\frac{m}{k} \ln \left(1 - \frac{t}{m} \right) \quad (1)$$

In addition to membership testing, Bloom filters also support union and intersection operations. An intersection of two sets \mathcal{D}_A and \mathcal{D}_B can be done by taking their respective Bloom filter representations and conducting a bitwise AND operation. To illustrate, if A is represented by $[0, 1, 1, 0, 1]$ and B by $[1, 1, 1, 0, 0]$, then $A \cap B$ is represented by $[0, 1, 1, 0, 0]$. A union is computed through a bitwise OR operation. (Note that for realistic representations of sets, Bloom filters generally have lengths of 1000s of bits.) Whereas unions do not affect the probability of false detections, intersections do. This also means that estimating the size of an intersection when using Bloom filters may easily see deviations. We ran extensive tests and encountered estimates that were 15% off the real size. A more accurate estimation for two intersecting sets is provided by [8], yet no general estimation is known for more than two intersecting sets. For this paper, we will use the simple approximation given by Equation 1.

Ignoring encryption for the moment, detections at a scanner s are converted into Bloom filters and sent by s to the server at the end of each epoch. To answer queries, the server may do a series of unions and intersections on various Bloom filters, as we explained in our simple

example above. The result is a Bloom filter representing the detections related to the query. At that point, the server could return the estimated cardinality of the set. Unfortunately, the server itself can still, with some computational effort, discover the detected card identifiers. As we mentioned, in our system model, we do not trust the server. This is where encryption comes into play.

3.2 Homomorphic Encryption

To prevent the server from discovering identifiers, Bloom filters must be combined with encryption schemes. Homomorphic encryption [9] is a specialized form of encryption that enables mathematical operations to be conducted directly on encrypted data without the need for decryption. The results of these operations are also encrypted, and the output is the same as if the operations had been conducted on unencrypted data.

The following procedure is now followed using homomorphic encryption. Suppose a user U is interested in knowing how many travelers moved from A to B . To that end, she passes an *encryption key* to the server, which is then used to encrypt all Bloom filters from the moment the key is available (note that this means that a user cannot issue queries that relate to the past, i.e., the time before they made the encryption key available). Also note that the user holds the *decryption key*, and is thus the only entity who can decrypt the corresponding encrypted Bloom filters. Neither the scanners nor the server can decrypt those Bloom filters.

The server now operates on bitwise encrypted Bloom filters and produces a final result, say an encrypted Bloom filter BF representing a set R . By simply *adding* the entries of BF , it can produce an (encrypted) version t^* of t , the number of bits that have been set to 1. This value, along with k and m can then be handed over to the user U , who can decrypt t^* and compute the cardinality c . The server can also hand out BF to the user, but not after having shuffled the entries (otherwise, the user could still decrypt BF and discover detections). Shuffling keeps the same number of (encrypted) bits that have been set to 1, but a shuffled version of BF has no relationship to R anymore.

4 Results and Discussion

In this section, to get a clear understanding of the effects of preserving privacy, we conduct an experiment by using a synthetic dataset. To determine the accuracy of the responses, we compare the statistical counts generated by our model with those from our dataset. For the hashing part, we choose MurmurHash3 [1], which is highly efficient. The estimation formula used by Bloom filters provides only an *approximation* of the number of elements likely to be present in the original set, rather than an exact count. For this reason alone, we expect to see deviations from the ground truth. In addition, taking intersections also affects the probability of having false positives; which will generally lead to overestimations of the size. We express the accuracy of the estimated count c to the real count c_t as:

$$Accuracy = \max\left(1 - \frac{|c - c_t|}{c_t}, 0\right) \quad (2)$$

To simulate real-world subway data, we generate card identifiers from a uniform distribution. As is common practice, real identifiers are often processed using a cryptographic hash function before being used for further analysis, and our use of uniform random identifiers similarly mimics this step. As an example, we ask ourselves how many travelers move from one station to another. Let $s_1^A, \dots, s_{n_A}^A$ be the sensors at station A and $s_1^B, \dots, s_{n_B}^B$ the sensors at station B, The answer is then $|\bigcup_{e_d} \bigcup_{e_a} \mathcal{D}_{e_d}^A \cap \mathcal{D}_{e_a}^B|$, where we assume that $e_d \triangleleft e_a$.

In other words, we take all combinations of departure epoch at A and *later* arrival epoch at B , and consider the detections from all sensors at A , and intersect that with the set of detections from all sensors at B .

To see the effects of taking intersections as unions, we count in two different ways. First, we simply compute the size of the union of intersections, as just mentioned. Second, we take a look at any combination of departure epoch e_d and (possible) arrival epoch e_a , as well as all pairs of sensors s_i^A and s_j^B . Using Bloom filter representations, we compute the size of the intersection $\mathcal{D}_{e_d}^A \cap \mathcal{D}_{e_a}^B$, and subsequently add those sizes for all combinations of departure and arrival epochs: $\sum_{e_d} \sum_{e_a} |\mathcal{D}_{e_d}^A \cap \mathcal{D}_{e_a}^B|$.

■ **Table 1** Comparison of the accuracy of estimated counts with ground truth.

Ground truth	100	1000	10000	100000
Estimated count first method	96	1006	10001	99933
Accuracy first method	96.00%	99.40%	99.99%	99.93%
Estimated count second method	97	990	10081	107670
Accuracy second method	97.00%	99.00%	99.19%	92.33%

We conducted four experiments using 100, 1000, 10000, and 100000 trips distributed over a single day (24h). For each experiment, we set the epoch length as 5 minutes (resulting in 288 epochs), fixed p at 0.001, and selected n to be equal to the corresponding number of trips in each experiment. We used optimal settings for m and k , given n and p . We ran each experiment 50 times. In both counting methods we perform a bitwise intersection with all possible arrival epochs for each departure epoch. The distinction between the counting methods takes into effect when performing intersections.

Table 1 presents the results of our experiments. The table displays the ground truth, as well as the estimated counts obtained using the two different counting methods from our proposed approach, along with the corresponding accuracy values. The results show that the difference between the counting methods becomes more pronounced as the number of trips increases. This is mainly because as epochs become more crowded, i.e., when we have more detections in a single epoch, the probability of false positives also increases when intersecting two epochs. The impact of false positives on the counting accuracy differs between the two methods. The first method yields an estimated count closer to the ground truth because it also considers the union of intersections. Taking the union ensures that false positives inside all intersections are counted only once because they are consolidated through the union operation at the end. In contrast, the second counting method estimates the size immediately after the intersection between each departure epoch at station A and each arrival epoch at station B. The estimated count after each intersection also includes false positives between the corresponding epochs. Therefore, the total count obtained at the end of this method is the sum of the counts obtained after each intersection, which includes false positives and leads to overestimating the total number of travelers. The first method's estimated count of travelers for all different numbers of trips is in close agreement with the actual count and consistently achieves high accuracy.

The difference in accuracies comes from the way we use Bloom filters, and is seen to be dependent on the query. Further research is needed to see how accuracies depend on different types of queries.

The current setup and implementation allow us to run queries involving (tens and hundreds of) thousands of travelers, often within just a few minutes, even with more intricate queries. When considering entire networks, many queries can be subdivided into independent parts, making them excellent candidates for processing in parallel.

5 Conclusion

In this paper, we have used a privacy-preserving method for counting travelers moving in public transport systems through encrypted Bloom filters. By using encrypted Bloom filters, we can count travelers moving between stations without revealing any information about who made these trips. Further, the information about who made which travels is unrecoverable and hidden for all components and parties in the system: the sensors, the server, and the client interested in the counts. The downside is that the method decreases the accuracy of counting. We evaluate the accuracy of our method on a synthetic subway dataset. We show that the loss of accuracy can be minimized and that it is possible to achieve highly accurate counting while ensuring that data cannot be used to trace back to an individual. An important observation is that the attainable accuracy is dependent on *how* counting takes place. If we count too soon to aggregate counts later on, we may fail to compensate for false counting later in the process. In other words, the accuracy of our method is dependent on the query and when counting and aggregation actually take place.

Although we did not show in this paper, our method is not limited to counting travelers moving between only two locations. The proposed method has the capability to handle more complex queries, such as counting the number of travelers moving between multiple locations. As a next step, we plan to investigate how more complex queries can also be answered with high accuracy. In addition, we need to investigate the practical feasibility of running queries such that answers can be provided in a reasonable time.


References

- 1 Austin Appleby. Murmurhash3.(2016). URL: <https://github.com/aappleby/smhasher/wiki/MurmurHash3>, 2016.
- 2 Kay W Axhausen, Andrea Zimmermann, Stefan Schönfelder, Guido Rindsfuser, and Thomas Haupt. Observing the rhythms of daily life: A six-week travel diary. *Transportation*, 29(2):95–124, 2002.
- 3 Burton H Bloom. Space/time trade-offs in hash coding with allowable errors. *Communications of the ACM*, 13(7):422–426, 1970.
- 4 Yves-Alexandre De Montjoye, César A Hidalgo, Michel Verleysen, and Vincent D Blondel. Unique in the crowd: The privacy bounds of human mobility. *Scientific reports*, 3(1):1–5, 2013.
- 5 Merkebe Getachew Demissie, Santi Phithakitnukoon, Titipat Sukhvibul, Francisco Antunes, Rui Gomes, and Carlos Bento. Inferring passenger travel demand to improve urban mobility in developing countries using cell phone data: a case study of senegal. *IEEE Transactions on intelligent transportation systems*, 17(9):2466–2478, 2016.
- 6 Yola Georgiadou, Rolf A de By, and Ourania Kounadi. Location privacy in the wake of the gdpr. *ISPRS International Journal of Geo-Information*, 8(3):157, 2019.
- 7 Dmytro Karamshuk, Chiara Boldrini, Marco Conti, and Andrea Passarella. Human mobility models for opportunistic networks. *IEEE Communications Magazine*, 49(12):157–165, 2011.
- 8 Odysseas Papapetrou, Wolf Siberski, and Wolfgang Nejdl. Cardinality estimation and dynamic length adaptation for bloom filters. *Distributed and Parallel Databases*, 28:119–156, 2010.
- 9 Ronald L Rivest, Len Adleman, Michael L Dertouzos, et al. On data banks and privacy homomorphisms. *Foundations of secure computation*, 4(11):169–180, 1978.
- 10 Valeriu-Daniel Stanciu, Maarten van Steen, Ciprian Dobre, and Andreas Peter. Privacy-preserving crowd-monitoring using bloom filters and homomorphic encryption. In *Proceedings of the 4th International Workshop on Edge Systems, Analytics and Networking*, pages 37–42, 2021.

A Personalised Pedestrian Navigation System

Urmi Shah¹ ✉ 🏠

School of Computing & Mathematical Sciences, University of Greenwich, UK
GeoLytix, London, UK

Jia Wang ✉ 🏠 

School of Computing & Mathematical Sciences, University of Greenwich, UK

Abstract

Many existing navigation systems facilitate pedestrian routing but lack the provision of personalised route alternatives tailored to individual needs. Previous research suggests that pedestrians often prioritise factors such as safety or accessibility over the shortest possible route. This paper investigates ways to enhance existing pedestrian navigation systems and improve walking experiences by providing personalised routes based on walking preferences. This is achieved by defining a set of routing preferences and implementing a modified version of Dijkstra's algorithm. The goal of this work is to promote walking by enhancing mobility, accessibility, comfort, and safety.

2012 ACM Subject Classification Information systems → Geographic information systems

Keywords and phrases Pedestrian, Navigation, Walking, Preference, Graph

Digital Object Identifier 10.4230/LIPIcs.GIScience.2023.67

Category Short Paper

1 Introduction

Currently, leading navigation applications do not (or rarely) provide pedestrians with personalised routes based on their needs. Among the widely used navigation apps utilised by pedestrians, notable examples include Google Maps, and Citymapper² which is particularly popular among commuters due to its emphasis on public transit. Both applications offer turn-by-turn directions from the starting point to the destination, with Citymapper performing better with live public transport information. It is important to note that both Citymapper and Google Maps predominantly suggest the shortest or fastest walking routes, and wheelchair-accessible routes are only available if public transport is incorporated into the journey (this is the case in London). The Mayor of London has launched the first ever Walking Action Plan³ for UK's capital city, and the mayor's vision is to make London the world's most walkable city by 2041. We aim to support this vision by implementing a pedestrian navigation application that provides personalised routes tailored to individual walking preferences, as opposed to solely recommending the fastest or shortest paths. A comparison with Google Maps shows that our generated routes excel in meeting users' walking needs by incorporating route features that align with user preferences.

2 Related Work

Many researchers have investigated the impact factors such as safety, travel purposes, weather conditions and traffic flow on pedestrian route choice [2, 7]. Bovy [2] reviewed theories and models of wayfinding behaviour and applied them in transport networks. He summarised

¹ Corresponding author

² <https://citymapper.com/london?lang=en>

³ <https://content.tfl.gov.uk/mts-walking-action-plan.pdf>



© Urmi Shah and Jia Wang;

licensed under Creative Commons License CC-BY 4.0

12th International Conference on Geographic Information Science (GIScience 2023).

Editors: Roger Beecham, Jed A. Long, Dianna Smith, Qunshan Zhao, and Sarah Wise; Article No. 67; pp. 67:1–67:6

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

the influential attributes into three categories, i.e., traveller, trip route, and circumstances. In Asha's study to learn routing choices [1], pedestrians cited safety and time-saving as the two most important factors when choosing their routes. One study determined seven criteria for pedestrian route choice: complexity, landmarks, accessible assistance, roadways, obstacles, intersections, and personal preferences [3]. In another study, the authors suggested SWEEP (Safety, Wealth, Effort, Exploration, and Pleasure) as significant route quality attributes utilised in a route recommendation system survey [10]. In recent years, there has been a growing interest in improving pedestrian navigation by integrating pavement facilities, the walking environment and pedestrian profiles for customised path finding [4, 11, 6, 8]. Fang *et al.*, [5] proposed a people-centric framework for pedestrian navigation based on three layers, namely physical sense, physiological safety, and mental satisfaction. The interdisciplinary review on mobile spatial navigation system [9] offered valuable design recommendations aimed at enhancing the accessibility and inclusivity of navigation systems. These recommendations encompassed the inclusion of physical accessibility information and the provision of personalised route options.

3 Pedestrian Routing Preferences

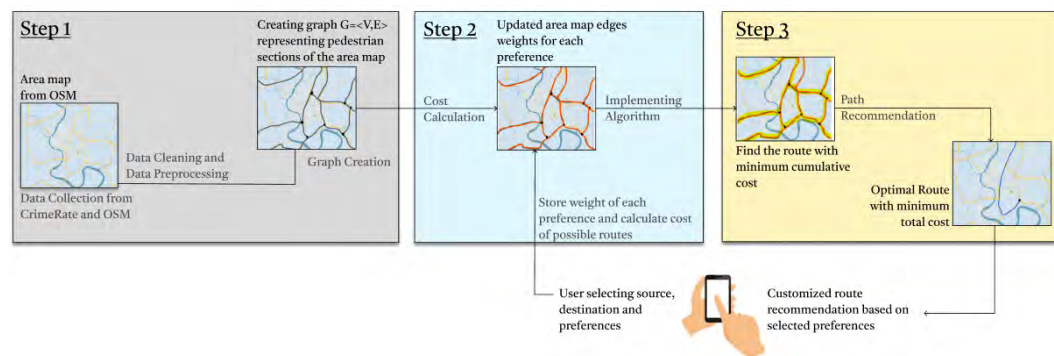
Routing preferences are quantified using a set of weights or costs, which are assigned based on characteristics related to the pavement and its surrounding environment. These preferences determine the type of route that will be chosen. Various studies have shed light on the primary determinant influencing pedestrians' route selection. For instance, one study [3] delves into the challenges confronted by visually impaired pedestrians when navigating between origin and destination, while another study [10] identifies a range of quality attributes, including safety and exploration, through a survey on route recommendations. Through an analysis of the existing literature, we have identified seven walking preferences: safety, presence of tactile paving, proximity to leisure areas, residential neighborhoods, low traffic volume, straightforwardness, and availability of step-free access.

- **Safety** preference indicates a secure and protected setting for pedestrians, which is particularly beneficial for individuals, especially women, who walk alone during nighttime. It is widely regarded as one of the most critical factors influencing pedestrians' route selection [9].
- **Tactile paving** enables individuals who are completely or partially blind to navigate along designated routes specifically equipped with tactile indicators.
- **Leisure spots** signifies routes with attractions such as green space, wetlands, shopping centres, and tourism spots which are particularly favoured by tourists and leisure walkers.
- The preference for a **residential neighborhood** highlights routes that primarily pass through residential areas, as opposed to industrial or commercial zones. This choice results in a quieter and less crowded path for pedestrians.
- The preference for **low traffic volume** prioritises pedestrians who prefer to avoid walking alongside high-volume motorways. This preference suggests routes with minimal traffic and reduced ground emissions and noises, providing a healthier and more pleasant walking experience.
- **Straightforwardness** promotes walking routes with minimal crossings and turns and is particularly targeted for joggers and elderly, or anyone who would typically prefer a straight path in order to avoid crossings and turns.

- The inclusion of **step-free access** suggests routes that are suitable for wheelchair users and accommodate the needs of individuals who require barrier-free access, e.g., those pushing pramchairs or carrying bulky luggage.

The street network is depicted as a graph, where street segments are represented by edges and street junctions are represented by nodes. Each edge in the graph has an initial weight value equivalent to its length. The suggested preferences are assigned to the edges as numerical values, which are then added to the base weight. Each edge is associated with a single weight value. The weight values are derived by pulling relevant data from OpenStreetMap (OSM)⁴ and CrimeRate⁵. OpenStreetMap uses a variety of tags to identify elements of street segments and junctions. Likewise, for each pedestrian preference, a corresponding OSM tag is defined, and its value is utilised to determine the weight assigned to a particular edge. The resulting route consists of a sequence of interconnected edges that link the source and the destination. The total cost of the route is determined by summing the weights of these edges. A modified version of Dijkstra's algorithm is employed to compute the costs of all the potential routes in such a way that the cost of each route is different based on the users' selection of preferences. While the conventional Dijkstra's algorithm examines all nodes in the graph to determine the shortest path between the starting and ending nodes, our approach employs Dijkstra's algorithm to analyse only specific nodes that are selected based on the users' preferences to identify the path with the lowest cost. The route with the lowest cost is deemed optimal, while the other routes are considered less favorable.

4 Prototype and Evaluation



■ **Figure 1** Workflow of the development of a prototype pedestrian navigation system.

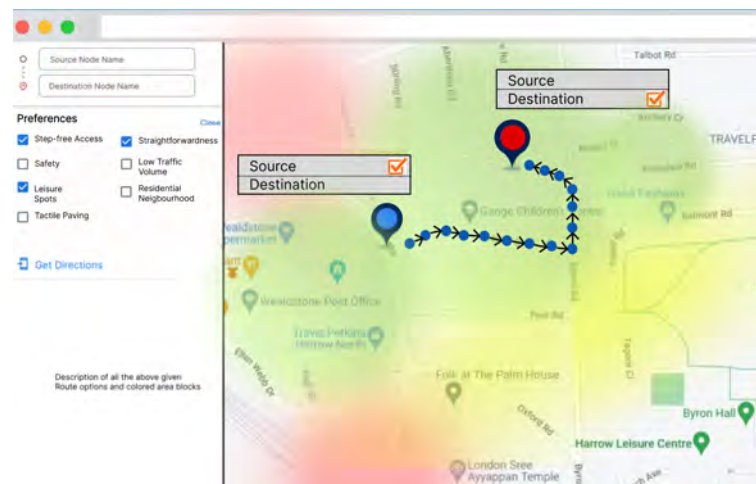
Figure 1 depicts a workflow flowchart of the prototype navigation system, with critical steps in each block represented. Two main data sources, OpenStreetMap and CrimeRate, are utilised in developing the prototype of a pedestrian navigation system. The case study focuses on an area measuring 0.7km² situated in the borough of Harrow in northwest London. The selection of this area is deliberate as it offers sufficient pavement and surrounding environment features that cover the proposed walking preferences. The OpenStreetMap data is used to construct the underlying geographical map for producing graphs that depict street networks. Crime incidents at the street level within the study area are collected from

⁴ <https://www.openstreetmap.org/>

⁵ <https://crimerate.co.uk/>

67:4 A Personalised Pedestrian Navigation System

CrimeRate between December 2019 and November 2022. This data source was collected to determine the weight value of “safety”. These crime incidents are classified into 12 distinct types of crime: burglary, damage & arson, drugs, other crime, other theft, possession of weapons, public order crimes, robbery, shoplifting, theft from person, vehicle based crimes, and violence and sexual crimes. This data will undergo data cleaning and preparation to convert it from its raw form (such as .osm/.xlsx) and to remove any incomplete data for further operations.



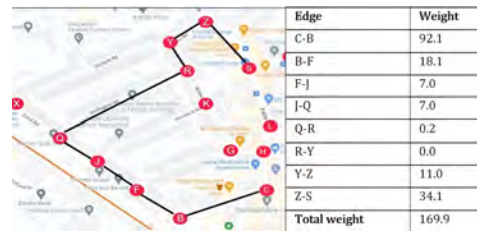
■ **Figure 2** A visualisation of the suggested route based on user preferences by the prototype.

Figure 2 shows the web-based prototype interface, with three required inputs as the source (blue pin), the destination (red pin), and three chosen preferences from the list of seven preferences displayed in the left panel. The user can either click on map or type in the source and destination input fields in the left panel to indicate where they want to go. According to Figure 2, A user has made a specific request for a route that fulfills the following criteria: step-free access, includes leisure spots, and has a minimal number of turns. The displayed route connecting the blue pin and the red pin is suggested by the algorithm in such a way that it meets the walking preferences. The coloured bubbles on the map indicate whether or not an area contains routes that meet the user’s preferences. Routes within the green bubbles are those that meet user preferences to the highest degree (lowest cost), while red represents routes that meet user preferences to the lowest degree (highest cost). A yellow highlighted area indicates that the routes in that area are moderately meeting user preferences (medium cost). The purpose of the bubble visualization is to enhance pedestrians’ awareness of their surroundings, enabling them to explore other route alternatives and at the same time avoid unpleasant walking experiences.

The evaluation of the prototype is carried out between the routes computed by the proposed routing algorithm (R_1) and those generated by Google Maps (R_2). In Figure 3, it can be seen that R_2 suggests a straight route from C to S , whereas R_1 in Figure 4 suggests a completely different route, which circles around the major nodes H and L and takes a longer journey to reach the destination. The reason for this is that according to CrimeRate, a higher number of reported crimes is reported in the vicinity of nodes H and L . Since safety is always the top priority, the prototype generates route R_1 , which bypasses these crime-prone areas thus is safer with a total cost of 169.9. In contrast, the route suggested by Google Maps does not avoid the crime points and has a higher total cost of 270.2.



■ **Figure 3** Suggested route from C to S by Google Maps (R_2).



■ **Figure 4** Suggested route from C to S by our method prioritising safety (R_1).

Figure 6 shows a route suggested by R_1 from processing multiple preferences (tactile paving and straightforwardness) selected by a user. Graham Road (X-Y-Z) and Grantt Road (N-O-P) are more accessible for visually impaired pedestrians as compared to the route suggested by R_2 suggested by Google Maps (Figure 5). Regarding the straightforwardness of the route, there are no crossings on the route suggested by R_1 and there is one crossing on the route R_2 suggested by Google Maps. Therefore, the most optimal route is R_1 with a total cost value of 17.8 compared to R_2 with a total cost value of 22.3.



■ **Figure 5** Suggested route from P to X by R_2 .



■ **Figure 6** Suggested route from P to X by R_1 with multiple preferences.

5 Conclusions and Future Work

This paper introduces an improved pedestrian navigation system that offers personalised routes considering seven walking preferences. The study primarily focuses on addressing the needs of commuters, particularly female pedestrians who have to traverse longer distances during night time. Additionally, the system aims to cater to pedestrians with physical challenges, the elderly, individuals accompanied by children or infants, leisure walkers and tourists. The implemented prototype is capable of computing a range of personalised routes, allowing users to select up to three preferences. This functionality goes beyond what major existing navigation applications currently offer. We believe the proposed system can encourage more walking by increasing confidence, safety and comfort in travel.

To enhance the proposed algorithm, it is suggested to incorporate open spaces, such as squares and parks, into the graph representation. The algorithm can be further optimised by addressing scenarios where conflicting preferences arise, such as situations where a route cannot simultaneously be the safest and the shortest. Enhancements can be made to ensure that the algorithm can provide more balanced routes that strike a suitable compromise between different preferences. Future work also includes taking into account temporal attributes when it comes to pedestrian routing, as time has a significant impact on walking needs. The initial application can be improved by incorporating adjustments that allow the display of information (such as walk time, reported crimes, etc.) within the colored bubbles. This enhancement will enable visual representation of route comparisons in a way that is easily understandable. Additionally, conducting a human study to evaluate the real user experiences of the prototype would be beneficial in assessing its effectiveness and gathering valuable feedback.

References

- 1 Asha Weinstein Agrawal, Marc Schlossberg, and Katja Irvin. How far, by which route and why? a spatial analysis of pedestrian preference. *Journal of Urban Design*, 13(1):81–98, 2008. doi:10.1080/13574800701804074.
- 2 Piet H Bovy and Eliahu Stern. *Route choice: Wayfinding in transport networks: Wayfinding in transport networks*, volume 9. Springer Science & Business Media, 2012.
- 3 Achituv Cohen and Sagi Dalyot. Route planning for blind pedestrians using openstreetmap. *Environment and Planning B: Urban Analytics and City Science*, 48(6):1511–1526, 2021. doi:10.1177/2399808320933907.
- 4 Ioannis Delikostidis, Corné P.J.M. van Elzakker, and Menno-Jan Kraak. Overcoming challenges in developing more usable pedestrian navigation systems. *Cartography and Geographic Information Science*, 43(3):189–207, 2016. doi:10.1080/15230406.2015.1031180.
- 5 Zhixiang Fang, Qingquan Li, and Shih-Lung Shaw. What about people in pedestrian navigation? *Geo spatial Inf. Sci.*, 18(4):135–150, 2015. doi:10.1080/10095020.2015.1126071.
- 6 Joan Henderson. Making cities more walkable for tourists: A view from singapore’s streets. *International Journal of Tourism Cities*, 4(3):285–297, 2018. doi:10.1108/IJTC-11-2017-0059.
- 7 Serge P Hoogendoorn and Piet HL Bovy. Pedestrian route-choice and activity scheduling theory and models. *Transportation Research Part B: Methodological*, 38(2):169–190, 2004. doi:10.1007/978-94-009-0633-4.
- 8 David Jonietz. Personalizing walkability: A concept for pedestrian needs profiling based on movement trajectories. In Tapani Sarjakoski, Maribel Yasmina Santos, and L. Tiina Sarjakoski, editors, *Geospatial Data in a Changing World - Selected Papers of the 19th AGILE Conference on Geographic Information Science, Helsinki, Finland, 14-17 June 2016*, Lecture Notes in Geoinformation and Cartography, pages 279–295. Springer, 2016. doi:10.1007/978-3-319-33783-8_16.
- 9 Ian T. Ruginski, Nicholas A. Giudice, Sarah H. Creem-Regehr, and Toru Ishikawa. Designing mobile spatial navigation systems from the user’s perspective: an interdisciplinary review. *Spatial Cogn. Comput.*, 22(1-2):1–29, 2022. doi:10.1080/13875868.2022.2053382.
- 10 Panote Siriaraya, Yuanyuan Wang, Yihong Zhang, Shoko Wakamiya, Péter Jeszenszky, Yukiko Kawai, and Adam Jatowt. Beyond the shortest route: A survey on quality-aware route navigation for pedestrians. *IEEE Access*, 8:135569–135590, 2020. doi:10.1109/ACCESS.2020.3011924.
- 11 Jia Wang, Zena Wood, and Michael F. Worboys. Conflict in pedestrian networks. In Tapani Sarjakoski, Maribel Yasmina Santos, and L. Tiina Sarjakoski, editors, *Geospatial Data in a Changing World - Selected Papers of the 19th AGILE Conference on Geographic Information Science, Helsinki, Finland, 14-17 June 2016*, Lecture Notes in Geoinformation and Cartography, pages 261–278. Springer, 2016. doi:10.1007/978-3-319-33783-8_15.

Estimating the Impact of a Flood Event on Property Value and Its Diminished Effect over Time

Nazia Ferdause Sodial ✉

City, University of London, UK

Oleksandr Galkin ✉

City, University of London, UK

Aidan Slingsby ✉

City, University of London, UK

Abstract

With the increase in natural disasters, flood events have become more frequent and severe calling for mortgage industries to take immediate steps to mitigate the financial risk posed by floods. This study looked more closely at the underlying effects of flood disasters on historical house prices as part of a climatic stress test. The discount applied on house prices due to a flood event was achieved by leveraging a causal inference approach supported by machine learning algorithms on repeat sales property and historic flood data. While the Average Treatment Effect (ATE) was employed to estimate the effect of a flood event on house prices in an area, the Conditional Average Treatment Effect (CATE) aided in overcoming the heterogeneous nature of the data by calculating the flood effect on property prices of each postcode. LightGBM as a base estimator of the causal model worked as an advantage to capture the nonlinear relationship between the features and the outcome variable and further allowed us to interpret the contribution of each feature towards the decay of these discounts using SHAP values.

2012 ACM Subject Classification Computing methodologies → Machine learning approaches

Keywords and phrases Flood, Causal Inference, Machine Learning, Property Analytics

Digital Object Identifier 10.4230/LIPIcs.GIScience.2023.68

Category Short Paper

Acknowledgements This research was conducted under the initiative of MIAC Analytics LTD. The flood data was acquired by MIAC Analytics LTD from its data partners, WhenFresh.

1 Introduction

Buying houses has been a constant activity despite the surge in property value. Mortgage industries have profited from these purchases with the rise in house prices. However, with the drastic climate change, there has been a concern about these changes reflected in household insurance policies and house prices. There are two major climate change risks: physical and transitional. Physical risks are the direct risks posed by the physical impact of climate change like global warming, ocean circulation change, high flood levels, etc. These risks represent losses brought on by the more frequent and severe hazards or events related to the environment. This introduces transitional risks, or those brought on by market, technical, legal, and policy changes resulting from the transition to a low-carbon economy. With 1.9 million people in the UK exposed to the river, coastal, or surface water flooding on a regular basis, this danger is already of a high scale and is expected to grow further in the absence of higher degrees of flood risk mitigation [4]. A lot of mortgage organizations developed



© Nazia Ferdause Sodial, Oleksandr Galkin, and Aidan Slingsby;
licensed under Creative Commons License CC-BY 4.0

12th International Conference on Geographic Information Science (GIScience 2023).

Editors: Roger Beecham, Jed A. Long, Dianna Smith, Qunshan Zhao, and Sarah Wise; Article No. 68; pp. 68:1–68:6

Leibniz International Proceedings in Informatics



LIPIC Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

68:2 Estimating the Impact of a Flood Event

an interest to participate in greening the financial system as a consequence of this drastic climate change as it could challenge the solvency of the companies. This has given rise to the demand for understanding the effect of flood events on house prices.

In order to better understand if severe climate change might have a significant impact on property values, it was attempted in this study to determine if historical flood events had an impact on house prices. The focus of this research was to identify a method that could handle the heterogeneity of the data and capture the non-linear relationship of the variables, which has been a challenge in this field of research. META Learner framework introduced in the `causalml` Python package was used to estimate the effect of the 2013 winter flood on Twickenham house prices and assess whether the effect eventually fades away with time.

2 Related work

Lamond *et al* [5] used coefficients of the regression to estimate the depression in the growth of house prices of part of the UK within a flood zone.

Beltrán *et al* [1] used a similar approach, modifying the repeat sales equation and using the coefficients to estimate the effect of flood events on different features of house prices. The findings indicate that for the majority of property types, the average post-flood price markdown of flood-affected properties is significant but generally transient.

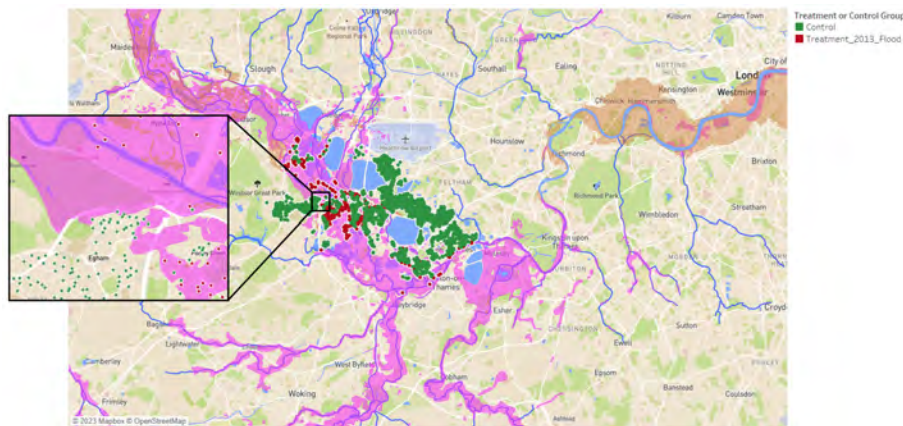
Again, N. Bui *et al* [2] used a similar approach by integrating “the hedonic property model in a difference-in-differences framework” and identified a discount of 9% was applied to house prices in some parts Ho Chi Minh City, Vietnam as a result of a disastrous flood event on 30 September 2017.

The aforementioned papers offer insightful and useful techniques into how flood occurrences affect home prices and the price decay that follows. In this study, causal models supported by machine learning algorithms were used in an effort to enhance the findings.

3 Methodology

3.1 Data

Hedonic approaches need many characteristics of each sold unit. Case and Shiller [3] recommended a different strategy that uses the information on units sold repeatedly. Repeat sales technique proponents contend that because it is based on the actual housing units’ observed appreciation, it more correctly accounts for property characteristics [3]. In this research, the repeat sales/transactional data were provided by MIAC Analytics Ltd which included previous price, recent price, previous transaction date, recent transaction date, property type, and the geographic details of the properties from 1995 to 2019. The flood data provided by the data provider of MIAC Analytics Ltd, WhenFresh Ltd included the history of all the flood events like flood cause, flood count, flood source, flood start date, flood end date, and property level details from the year 1900 to 2020. The elevation of each postcode was gathered from Ordnance Survey, and Shapefiles of the historic flood map, river and coastal bodies of the UK, area benefiting from flood defense, and geography level of the UK were collected from DEFRA and EA. This study was conducted on the data from 2010 to 2020 to avoid the impact of the house price crash of 2008.



■ **Figure 1** Map of Twickenham. Blue lines: river bodies; orange shading: areas protected by flood defence; pink shading: historic flood shadings; red shading: overlap of historic flood shading and areas protected by flood defence; red dots: postcodes only impacted by 2013 flood; green dots: postcodes which did not experience flood events and are outside flood risk areas.

3.2 Assumptions

When a property's structural attributes remain constant between transactions, one or more of the following variables could be to blame for the price variation: inflation, significant local changes like the construction of new transportation infrastructure or a flood occurrence, and random variation [5]. It is vital to make the assumption that all properties are equally affected by changes in location variables other than flood occurrences when building repeat sales models. By selecting relatively small areas for investigation and by getting access to local knowledge about any significant events, this can be made more likely [5]. Due to the above assumptions, all the analysis on flood effects were conducted on smaller regions that are geographically adjacent to each other. The districts TW15, TW16, TW17, TW18, TW19, and TW20 of Twickenham were considered for this analysis as the areas were partially covered in historic flood, flood risk, and flood defence. Every other flood-affected region of the UK either had fewer data points or is well protected by flood defences. The 2013 winter flood had an influence on 1404 Twickenham transactions. Thus, Twickenham and the 2013 flood disaster were taken into account for the subsequent analysis.

3.3 Treatment Control Group

The winter flood of 2013 was considered as the treatment effect in this research. While a small group of districts in Twickenham was considered in this study to hold other effects on the house price constant, the area was further divided into treatment and control groups. The treatment group was created as a collection of postcodes that experienced the 2013 winter flood, with the flooding having no influence on transactions prior to treatment but having an impact on transactions after treatment. The control group had the collection of postcodes that were never impacted by the flood and are outside flood risk as designated by the Environmental Agency.

3.4 META Learners

META Learner is a framework that uses multiple base learners/machine learning algorithms to build a model to estimate Average Treatment Effect (ATE) and Conditional Treatment Effect (CATE). ATE is the effect of a treatment on a population whereas CATE is the effect

of a treatment on a subgroup of the population based on a condition. `causalml` [6] is a Python package provides the tree models – Random Forest, LightGBM, and XGBoost – as the base learners for the META Learners. In this research, the four types of META Learners: S, T, X, and R Learners were used to estimate and validate CATE. To validate which base learner would perform the best in estimating the treatment effect, the preprocessed covariates and outcome variables were passed through all three algorithms, and random search hyperparameter tuning was applied to determine the best parameter. Once the best machine learning algorithm was selected as the base algorithm, the data were passed through all the META learners to calculate CATE. The average of CATE was used to calculate Average Treatment Effect (ATE) which indicated the flood effect on Twickenham.

3.5 Equations

Lamond *et al* [5] explain the derivation of the equations used to estimate the treatment effect. This approach was used to determine the growth effect by considering the market effect to be constant. For property i , the growth in price (P) from time t to $t + k$ is:

$$Y = \ln(P_{i(t+k)}) / \ln P_{it}$$

This term was used as the outcome variable to estimate the flood effect on the house prices of Twickenham. While each META Learner (S, T, X, and R) has different equations, in this section the equation of T learner is discussed as it proved to be the best learner. In the first stage, the T learner estimates the average outcome using machine learning algorithm[6]:

$$\mu_1(x) = E[Y(1)|X = x]$$

$$\mu_0(x) = E[Y(0)|X = x]$$

where μ is the average outcome, 0 indicates the non-flooded (control) properties, 1 indicates the 2013 winter flood-impacted (treatment) properties, X values are the features, and Y is the outcome variable or growth effect. In the second stage, CATE ($\hat{\tau}$) is estimated using the below equation[6]:

$$\hat{\tau}(x) = \hat{\mu}_1(x) - \hat{\mu}_0(x)$$

The ATE was further estimated by calculating the average of CATE.

4 Results and discussion

Although XGBoost outperformed all the models slightly, LightGBM was chosen as the base learner as it captured the contribution of most of the features which seemed more ideal to estimate the flood effect with an RMSE of 0.23. While ATE was efficient in capturing the impact of the 2013 winter flood on Twickenham districts, CATE was effective in capturing the impact of the flood event on each Twickenham postcode. With LightGBM as the base learner, the results of the META learners are below:

■ **Table 1** The flood effect estimated by all the META Learners of 2013 winter flood in Twickenham districts. The negative sign indicates that the flood effect resulted in a discount on house prices.

META Learners	S	T	X	R
ATE	-0.07	-0.08	-0.11	-0.07

4.1 Validation

The `causalml` package [6] was built using synthetic data. To validate the learners, the actual treatment effect was generated from the synthetic data and then the values of ITE and the actual treatment effect were used to validate the results and chose the best learner. However, in real-world applications, since the ground truth is not available, choosing the best learner was quite challenging. So, cumulative gain plots with a theoretical curve produced by the random model were used to validate the performance of the META learners. Once the theoretical random curve was in place, then the learners were compared to it as a benchmark. Every curve had the same beginning and end. Since the curve of the T learner deviated the most from the random line, it was considered to be the best learner amongst S, X, and R learners and the ATE value of the T learner was statistically significant. Meanwhile, from the permutation importance, it was observed that the T learner was capturing the effect of almost all the features. Hence, it was concluded that the T learner estimated the most authentic flood effect, and a discount of 8% was applied to the districts of Twickenham in 2019 as a result of the 2013 winter flood.

4.2 Diminishing of flood effect over time

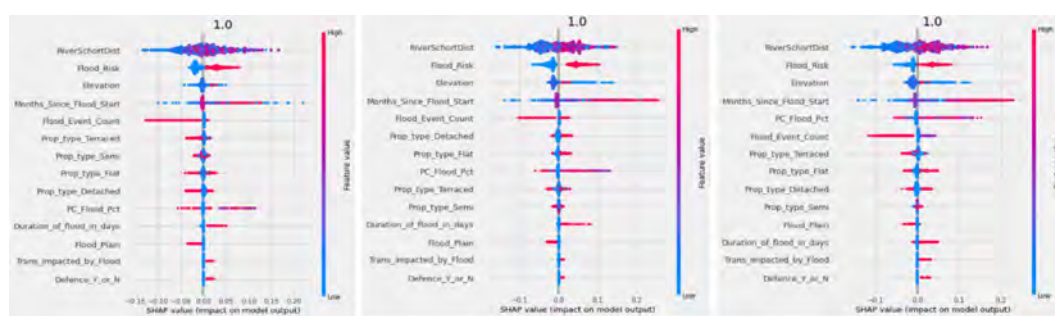
SHAP (Shapley Additive Explanations) values were used to interpret the Meta Learners. Each feature is given an importance value by SHAP for a specific prediction. These values were used to understand the Meta Learners better. As years pass by, customers tend to forget about the flood event, and in the absence of any other issues with the property, house prices tend to continue to increase. A slight decay in discount after 4 years of the flood event can be observed in Table 2. Fig. 2 shows that as the months between the transaction and the flood increase, there is growth in house prices. Whilst for semi-detached houses the discount remained the same from 2015 to 2019, for flats, terraced, and detached houses the discount decays and data points tend to contribute towards the growth of house prices by 2019. The SHAP plots also indicate that lesser values of the features: (a) distance from a river body, (b) elevation from mean sea level, (c) months since the flood event happened to contribute to higher flood discount. While properties within flood plains that experienced a higher number of flood events contribute to flood discounts, the areas protected by flood defence contribute to the growth of house prices.

■ **Table 2** The flood effect estimated by T learners with CI of Twickenham districts over the years.

Years after 2013 flood	Lower Bound	ATE	Upper Bound
2 years	-0.16	-0.09	-0.02
4 years	-0.16	-0.09	-0.02
6 years	-0.16	-0.08	-0.02

5 Conclusion

If flood effects do not reflect on the house prices, it raises the concern that a sudden risk re-pricing could be financially unstable if this risk is not represented in house prices. The two major papers that dealt with a similar goal [5, 1] used the "repeat sales method" along with linear or generalized regression method to determine the discount. This study contributes to the ongoing research in the field of climate change and its impact on transitional risk. The use of the causal inference algorithm backed by machine learning, Meta learners, presented



■ **Figure 2** The SHAP plot of T learner over the years. *Left*: after 2 years of the flood event (2015). *Middle*: after 4 years of the flood event (2017); *Right*: after 6 years of the flood event (2019).

a significant answer to the problems with the previous research efforts. First, by using CATE to estimate the flood effect on each postcode, it is possible to capture heterogeneity. Secondly, capturing the non-linear correlations between the data using LightGBM. Third, the algorithms' ability to be understood by using SHAP values.

While Beltrán *et al* [1] stated that “the discount is short-lived and the discount is no longer statistically significant for properties affected by inland flooding after 5 years, which falls to just 4 years for properties affected by coastal flooding”, we found that the discount begins to diminish after 4 years following the 2013 flood event. However, with a longer timeline, it could have been more interesting to capture the decay in the flood discount. One of the shortcomings of this project would be the fact that the treatment's random assignment is not assured as a lot of factors could contribute to the flood occurrences. Beltrán *et al* [1] agreed, as flooding occurs mostly in areas that are exposed to flood risk/hazard. It could be argued that the property or real estate market in such areas might already possess some special characteristics and attract households with distinctive preferences. The research was unable to calculate the treatment effect independently for the four property types due to insufficient data as a result of concentrating on a narrower area to hold other effects on the house prices constant. Although the results of the treatment effects cannot be generalized due to the assumptions associated with the repeat sales data, this methodology can be used to estimate the effect of flood events on house prices for any area and any timeline.

References

- 1 Allan Beltrán, David Maddison, and Robert Elliott. The impact of flooding on property prices: A repeat-sales approach. *Journal of Environmental Economics and Management*, 95:62–86, 2019.
- 2 Nam Bui, Le Wen, and Basil Sharp. House prices and flood risk exposure: An integration of hedonic property model and spatial econometric analysis. *The Journal of Real Estate Finance and Economics*, pages 1–32, 2022.
- 3 Karl E Case and Robert J Shiller. Prices of single family homes since 1970: New indexes for four cities, 1987.
- 4 CCRA3. Housing briefing. <https://www.ukclimaterisk.org/wp-content/uploads/2021/06/CCRA3-Briefing-Housing.pdf>, 2021. [Online; accessed 17-April-2023].
- 5 Jessica Lamond, David Proverbs, and Adarkwah Antwi. Measuring the impact of flooding on uk house prices: A new framework for small sample problems. *Property Management*, 2007.
- 6 Uber. Meta-learner algorithms. <https://causalml.readthedocs.io/en/latest/methodology.html>, 2022. [Online; accessed 06-Nov-2022].


Development and Operationalisation of Local Sustainability Indicators - A Global South Perspective on Data Challenges and Opportunities for GIScience

Stefan Steiniger  

Pontificia Universidad Católica de Valparaíso, Valparaíso, Chile
CEDEUS, Santiago, Chile

Carolina Rojas 

Pontificia Universidad Católica de Chile, Santiago, Chile
CEDEUS, Santiago, Chile

Ricardo Truffello 

Pontificia Universidad Católica Chile, Santiago, Chile
CEDEUS, Santiago, Chile

Jonathan Barton 

Pontificia Universidad Católica de Chile, Santiago, Chile
CEDEUS, Santiago, Chile

Abstract

Evaluating and monitoring the sustainable development of nations and cities requires sets of indicators. Such indicator sets should measure equity, health, environmental, or governmental progress or recess - among other sustainability aspects. In 2015 the United Nations ratified 17 Sustainable Development Goals (SDG) assessed through 231 indicators. However, other - local - sets of indicators have been developed too. In this paper we review geodata challenges that emerged when we developed four sustainability indicator sets in Chile. Faced challenges include (geo)data availability and data representativeness, among others. We analyse how GIScience knowledge has contributed to indicator development and outline three priority research topics: (i) updating indicators based on automated processes, while respecting representativeness, (ii) tools for planning scenario generation, and (iii) methods for short- and long-term forecasting.

2012 ACM Subject Classification Information systems → Geographic information systems

Keywords and phrases geographic information, SDGs, indicators, sustainable development, Chile

Digital Object Identifier 10.4230/LIPIcs.GIScience.2023.69

Category Short Paper

Funding The authors acknowledge funding from CEDEUS – Centro de Desarrollo Urbano Sustentable (ANID/Fondap/1522A0002).

1 Introduction

In 2015 the United Nations adopted a new development agenda with the title “Transforming our world: the 2030 Agenda for Sustainable Development”. In this agenda 17 Sustainable Development Goals (SDG) and 169 particular development targets are outlined that should be met until the year 2030. The first 4 sustainable development goals address people’s basic demands: (1) no poverty, (2) zero hunger, (3) good health and wellbeing, and (4) quality education. Further important goals include gender equality, clean water, responsible consumption, as well as reduced inequalities. Given that today 54 percent of the global population lives in cities, Goal 11 has been targeted at cities. This goal has a focus on



© Stefan Steiniger, Carolina Rojas, Ricardo Truffello, and Jonathan Barton;
licensed under Creative Commons License CC-BY 4.0

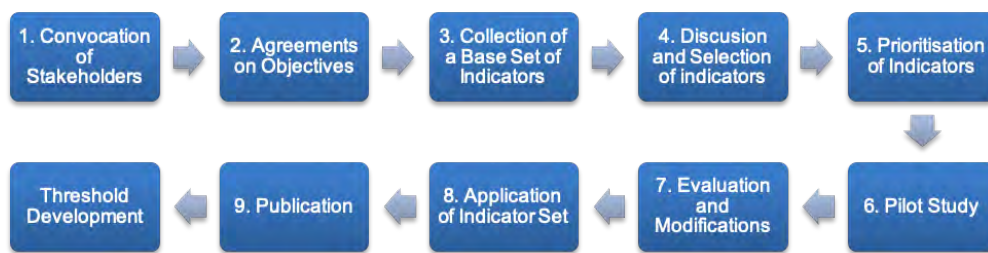
12th International Conference on Geographic Information Science (GIScience 2023).

Editors: Roger Beecham, Jed A. Long, Dianna Smith, Qunshan Zhao, and Sarah Wise; Article No. 69; pp. 69:1–69:6

Leibniz International Proceedings in Informatics



LIPIC Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany



■ **Figure 1** Typical process to develop a set of urban sustainability indicators.

“Make[ing] cities and human settlements inclusive, safe, resilient and sustainable”. To evaluate progress towards reaching the development targets the SGDs are accompanied by a set of indicators, currently being 231. However, not only the UN has published indicators that measure coverage of people’s basic needs, sustainability, and quality of life aspects. Other well-known globally applied indicator sets are for instance the Human Development Index and the Environmental Performance Index with a focus on assessing and comparing nations. But there are also indicator sets with a focus on cities, such the Happy Planet Index and the Global City Indicators (see [12]). For Chile, home country of the authors, several indicator sets have been developed as well. The authors directed and participated in the development of at least four sustainability indicator sets, including the CEDEUS Indicators (<http://indicadores.cedeus.cl>; [12]) and the SIEDU indicators published by the National Council for Urban Development (CNDU). Given these experiences on indicator development and operation, we wanted to review our work guided by the following question: “*What have been the difficulties during the operationalisation of urban indicators from a geoinformation perspective and how did GI experts contribute?*”. To answer this question, we outline first the process to develop indicators (Section 2) and then summarize the (geo)data issues that we faced (Section 3). We then reflect on our experiences by looking at the geoinformation team contributions (Section 4) and further research needed to advance in our work (Section 5).

2 The development process of urban sustainability indicator sets

To better understand the context of difficulties that may be experienced during the development of indicators sets, we will outline the steps that may be used. We identified 9 steps that can be presented as consecutive steps (see Figure 1). However, in reality the development process often includes several iterations until consensus among the stakeholders may be reached. (1) The first step of the indicator development process usually addresses the convocation of stakeholders that may be interested in using the indicators later on. Stakeholders may include experts, including researchers, representatives from municipal and governmental administration, local civic organizations, and NGOs [1]. (2) The first meetings with these stakeholders have the objective to identify the purpose of the indicator set and finding a common language. (3) In the third step a base collection of indicators may be developed usually through an analysis of the literature with input from the stakeholders. Often the resulting base set consists of several hundred indicators. (4) In the following step the indicators of this base set are then analysed and discussed to select indicators that can meet the earlier defined objectives. (5) As the resulting set of indicators may still be large (e.g. around 100 indicators), a prioritisation exercise may be carried out. (6) Next step is a pilot study that operationalises the indicators that are considered as of high priority. The



■ **Figure 2** Data related challenges experienced during indicator development.

pilot study serves several purposes, including for instance the selection of indicator variables and analysis of the calculation results so as to confirm that the indicator variables are able to highlight differences and trends among study sites. This phase is often executed by a geo-information or statistics team consulting with domain experts. (7) The results of the pilot study are then presented to all stakeholders for further discussion, so as to identify if the selected indicators and variables are able to fulfil their purpose, being sufficiently sensitive and robust. (8) Given that all stakeholders agree on the indicators and variables, the indicators are calculated for all cities or region(s) of interest. (9) Finally, indicator results are to be published to inform the public. While these are the basic steps of indicator development, it is possible that these are followed by an additional process with the objective to develop sustainability (and quality of life) thresholds for each indicator.

3 Geodata challenges

Given our experiences, in general the processes of developing indicator sets face the problem of data availability when indicators are proposed. While this may perhaps not be a surprise, if one considers that our work has a focus on Chile, a country in South America, the very same issue of data availability has also been highlighted by the authors of the U.S. Sustainable Cities Report in their 2017 version [10]. Prakash and colleagues considered to calculate values for 49 indicators for the 150 most populous “cities” originally, but had to resort to analyse only the 100 most populous “U.S. Metropolitan Statistical Areas (MSA)” - out of a total of 382 MSAs - due to a lack of and problems with data [10]. Besides the in-existence of data, a range of difficulties related to (geo)data can be found (Figure 2). In some cases data may exist, but it may be difficult to obtain access to it. Reasons for access issues are for instance concerns and confidentiality classifications by public administration, such as tax and crime records in Chile, or because the data owner is actually a state-contracted survey company or a public service provider, such as a private electricity and water services provider. Whereas access to these locked-up data may not be impossible for the purpose of indicator calculation eventually, it still raises the issue of completeness, transparency, and reproducibility for users of the indicators. Further difficulties arise from the fact that responsibilities for (public) urban data are often distributed among different levels of government (ministries, regions, cities), and public & private service companies (see also [4]). Result of these dispersed responsibilities is that data may lack complete geographical coverage and are dispersed as well. This meant for some of the CEDEUS indicators that data had to be requested from 71 municipalities.

Not few times requested (tabular) data may come printed, on CD as pdf, via (snail) mail or email, since (Spatial) data infrastructures exist only in some ministerial divisions and a hand full of municipalities - mostly due to a lack of experts and high cost of such infrastructure. Even if a geodata infrastructure exists, access is often possible only for in-house users and access via data APIs are rare.

Completeness issues do not only concern geographical coverage of data. It includes also temporal coverage and statistical representativeness of survey data. Temporal coverage, i.e. survey frequency, turns out to be an issue when cities grow rapidly, as in many developing nations, when population census surveys are performed only every 10 years. In this case evaluation and (city) planning is often based on outdated data and “monitoring” of change is complicated. Similarly, comparability among cities will be difficult if surveys are made in different seasons or years. These very same difficulties were already reported by Hoornweg et al. [5] 15 years ago (in 2007) for indicator projects at the World Bank, who found “a lack of reliable disaggregated data that are comparable across cities and over time.” In the case of Chile, for instance, the origin-destination travel surveys used for mobility indicators are carried out only for mayor cities every 10 years, and surveyed in different years.

A lack of statistical representativeness of survey data for the municipal level has posed as well problems when we aimed at operationalising indicators. Some important surveys in Chile, such as the two-yearly socio-economic household survey CASEN, are carried out to obtain statistics at regional level only, and are therefore representative only at this level. However, city planning requires data at least at municipal level.

4 GIScience contributions

Changing the perspective from reviewing the encountered data challenges for indicator implementation to the perspective of how GIScience knowledge & technology can facilitate the process of indicator operationalisation, calculation, and communication, we identify three broader areas of contribution:

Expertise on geodata - The expert team that is usually in charge of gathering data and calculating the indicators often contain geographers and statisticians. Geographic information experts are able to contribute here with their expertise when searching for indicator base data and in the assessment of the suitability of candidate datasets. If no data could be found that fits indicator requirements, such as geographical coverage, yearly data updates, and geographical and socio-demographic representativeness [12], then GI experts can help to establish criteria for the collection of new data. This includes to identify at what geographic scale data is needed, and what data collection tools, methods, and sampling schemes may be used to ensure representativeness.

GI Systems & Standards - A second area of GIScience contributions to indicator development and operation concerns the utilization of GI technologies and standards. To mention here are technical developments and standards related to Spatial Data Infrastructures (SDIs) [8]. The Open GeoSpatial Consortiums’ (OGC) standards for data description, cataloguing, and search help to identify suitable data sources for consecutive monitoring of urban conditions. Also other OGC standards, such as the OGC Simple Features specification implemented in spatial databases and the W*S specifications are essential to manage efficiently city or countrywide datasets. Finally, the OGC sensor web related standards permit updating sensor-based indicators that for instance assess environmental pollution or usage of transport modes. Other GI related tools have helped too: most prominently the scripting languages and processing frameworks that permit to automate data processing and analysis, such as Python and R.

The GI expert toolbox for indicator interpretation - There is a further area of expertise and opportunity for contribution that concerns the analysis of spatial data and a profound knowledge of the geostatistical analysis toolbox. Good indicator variables allow to identify differences and trends among cities or perhaps even neighborhoods. To assess a variables

geographical sensitivity that permits to identify where a policy intervention may be necessary, it is necessary to assess a variables distribution function and employ (geo-)statistical tests to identify the significance of trends and differences. The geostatistical toolbox employed for identifying meaningful indicator variables can help as well to define sustainability thresholds for indicators. Geospatial visualization tools are further useful in producing maps or (carto-)diagrams, during indicator development and when interpreting results. Maps like visualizations allow to validate visually an indicator variable and its data for coherence (with expectations) as well as its geographical sensitivity. For instance, energy consumption patterns that reflect income segregation between city neighbourhoods [12]. Visualization of indicators through maps and urban dashboards supports communication of indicator results to the public and decision makers [6].

5 Three emerging research topics for GIScience

Considering our four indicator development experiences we analysed what challenges are ahead of us when maintaining and utilizing the indicator sets. We identified three broader challenges:

Indicator updates from Sensor Data – The first challenge concerns the ability to monitor urban change; be it as a result of changes in natural (e.g. climate) or political conditions (e.g. new policies). For most indicators it is sufficient to update data and indicators only once a year, however, for some indicators monthly or even daily updates are possible. The challenge is here to update data and indicators ideally in an automatic fashion, be it from sensor data or civil service registries. This requires (widespread) introduction of data APIs and the implementation of data processing chains. Implementing these becomes challenging if one considers that the data need to be collected, evaluated, cleaned, and processed for different cities and regions in a way that always ensures plausible and representative indicator results, even if sensors fail or received data contain somewhat obscure values. Similarly, scaling up of indicator results from neighborhoods to city and to regional levels requires to account for maximal permissible errors.

Planning Scenario Modelling - A further need that we see concerns the development of a toolbox that permits to generate and evaluate city planning scenarios, as shown for instance by the AURIN project and the Urban FootPrint platform [9, 7]. We imagine it to be somewhat like an online version of the computer game “Sim City”, but rather with a focus on city policy tools & models, than the provision of a city construction toolset. The scenarios that are created are then to be evaluated with indicator sets to assess the impact of policy changes [11]. Even more interesting could be to develop indicator-based scenarios via back-casting, i.e. defining what indicator values need to be obtained in the future and see what needs to be done to get there [2].

Short- and long-term forecasting - To evaluate how indicator values may develop in the future, spatial models need to be developed for forecasting. Of interest may be indicator forecasting for a few days only, similar to meteorological forecasting, and forecasting for the next year or the next five years. While it is safe to assume that policies do not change when forecasting for a few days or weeks, forecasting for one year may and for 5 years actually should be able to consider policy changes, so as to explore impacts of policy changes. The task is here to keep working on forecasting methods that explicitly allow to include spatial interdependencies among indicators (see for instance Fotheringham et al. [3]).

6 Conclusions

As we have outlined, a paramount challenge for the development of indicator sets for the assessment of urban sustainability is that often required data are either not existent or not accessible. This includes in particular data needed to evaluate the UN's 17 Sustainable Development Goals (SDGs). Work by Prakash et al. [10] and ours show that the lack of data is a global problem. Hence, the expertise of GIScientists and GI professionals is needed to identify and collect required (geo)data.

While being able to contribute with knowledge on data and tools to indicator development and monitoring, we think that GIScientists need to promote the (still new) spatio-temporal perspective - overcoming a statistical and national perspective. This will allow to develop geographical targeted indicators and policies that may be more suited for geographically diverse countries. Similarly, we think that GIScience needs to promote cost and resources effective governance and responsibility models concerning data. We believe that Europe's INSPIRE directive and the European Environment Agency can lead as an example that shows how Spatial Data Infrastructures (SDIs) can facilitate access to data.

References

- 1 S. Amoushahi, A. Salmanmahiny, H. Moradi, A. R. M. Tabrizi, and C. Galán. Localizing sustainable urban development (sud): Application of an fdm-ahp approach for prioritizing urban sustainability indicators in iran provinces. *Sustainable Cities and Society*, 77:103592, 2022.
- 2 R. Crespo and A. Rajabifard. Inverse model using land and property sub-systems for planning future cities: A general framework. *Journal of Urban & Regional Analysis*, 14(1), 2022.
- 3 A. S. Fotheringham, R. Crespo, and J. Yao. Exploring, modelling and predicting spatiotemporal variations in house prices. *The Annals of Regional Science*, 54:417–436, 2015.
- 4 M. S. Fox and C. J. Pettit. On the completeness of open city data for measuring city indicators. In *2015 IEEE First International Smart Cities Conference (ISC2)*, pages 1–6. IEEE, 2015.
- 5 D. Hoornweg, F. Ruiz Nuñez, M. Freire, N. Palugyai, M. Villaveces, and E. W. Herrera. City indicators: Now to nanjing, 2007.
- 6 R. Kitchin, S. Maalsen, and G. McArdle. The praxis and politics of building urban dashboards. *Geoforum*, 77:93–101, 2016.
- 7 E. Pajares, B. Büttner, U. Jehle, A. Nichols, and G. Wulfhorst. Accessibility by proximity: Addressing the lack of interactive accessibility instruments for active mobility. *Journal of Transport Geography*, 93:103080, 2021.
- 8 G. Percivall. Progress in OGC web services interoperability development. In L. Di and H.K. Ramapriyan, editors, *Standard-Based Data and Information Systems for Earth Observation*, pages 37–61. Springer, Berlin, 2010.
- 9 C. J. Pettit, R. E. Klosterman, M. Nino-Ruiz, I. Widjaja, P. Russo, M. Tomko, R. Sinnott, and R. Stimson. The Online What if? Planning Support System. In S. Geertman, F. Toppen, and J. Stillwell, editors, *Planning Support Systems for Sustainable Urban Development*, LNGC, pages 349–362. Springer Berlin, 2013.
- 10 M. Prakash, K. Teksoz, J. Espey, J. Sachs, M. Shank, and G. Schmidt-Traub. The U.S. Cities Sustainable Development Goals Index 2017 - Achiving a sustainable urban America. Technical report, Sustainable Development Solutions Network, New York, NY, USA, 2017.
- 11 S. Steiniger, M. E. Poorazizi, and A. J. S. Hunter. Planning with citizens: Implementation of an e-planning platform and analysis of research needs. *Urban Planning*, 1(2):46–64, 2016.
- 12 S. Steiniger, E. Wagemann, F. de la Barrera, M. Molinos-Senante, R. Villegas, H. de la Fuente, A. Vives, G. Arce, J.C. Herrera, J.A. Carrasco, et al. Localising urban sustainability indicators: The cedus indicator set, and lessons from an expert-driven process. *Cities*, 101:102683, 2020.

Assessing Epidemic Spreading Potential with Encounter Network

Behnam Tahmasbi ✉

Dept of Civil and Environmental Engineering, University of Maryland, College Park, MD, USA

Farnoosh Roozkhosh ✉

Department of Geography, University of Georgia, Athens, GA, USA

X. Angela Yao ✉

Department of Geography, University of Georgia, Athens, GA, USA

Abstract

Densely populated urban public transportation systems can provide conducive environments for transmitting viruses via close human contact or touching contaminated surfaces. In network analysis, Betweenness Centrality (BC) has been used as the primary metric to measure a node's communication with others. This research extends from the concept of BC and develops new measures to assess the risk of transmitting disease through public transportation links. Three new concepts are introduced: source Total Betweenness centrality (TBC), target TBC, and Encounter Network. From a network node (source node), the set of shortest paths from that node to all other nodes composes a sub-graph (tree). The source TBC of this node is defined as the sum of BC of all edges of this tree. Similarly, using the shortest path tree consists of the set of the shortest paths from all nodes to the node as the destination, the target TBC of the node is defined as the sum of BC of all edges of this tree. Both TBC can be weighted by edge characteristics such as travel time or trip volume. Another new concept, Encounter Network, is constructed as the intersection between all source-target pairs of the public transportation network. We use the source TBC of a node to evaluate the relative risk of transmitting the disease from that node to other nodes. In contrast, the target TBC of a node can be used to assess the relative risk of being infected by a virus transmitted from other nodes to that node. A preliminary case study is conducted to illustrate the process and results.

2012 ACM Subject Classification Information systems → Geographic information systems; Networks; Networks → Metropolitan area networks; Applied computing → Transportation

Keywords and phrases Encounter Network, Total Betweenness Centrality, Complex Network, Epidemic spreading, Transmission risk, Public Transportation

Digital Object Identifier 10.4230/LIPICs.GIScience.2023.70

Category Short Paper

1 Introduction

Public transportation plays an essential role in many cities to achieve equitable and sustainable goals in urban systems [11]. Public transportation has been recommended in recent decades to reduce car dependency and externalities like traffic congestion and air pollution [2]. Despite its many positive contributions, mass transit network also provides a conducive environment for human contact in proximity which may lead to other effects. For instance, infectious diseases can be spread by human contact, especially in an enclosed space. Human contact between passengers in mass transit systems can easily facilitate the spreading of infectious diseases [7]. The crowded indoor environment in trains and buses intensifies the transmission of pathogens from infected passengers to others [10, 14]. To identify the network components where high-intensity of involuntary human contact and transmission may occur, this research proposes a new concept, encounter network (EN), a subgraph of a transit network, as well as related measures and algorithms to derive an EN from the transit network.



© Behnam Tahmasbi, Farnoosh Roozkhosh, and X. Angela Yao;
licensed under Creative Commons License CC-BY 4.0

12th International Conference on Geographic Information Science (GIScience 2023).

Editors: Roger Beecham, Jed A. Long, Dianna Smith, Qunshan Zhao, and Sarah Wise; Article No. 70; pp. 70:1–70:6

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Network science techniques, such as connectivity and centrality measures, have been widely used to study public transportation systems. The approach is useful for simplifying the transportation network by studying the network's topological properties [3], and it is also effective in studying and assessing the changes and modifications in the network and operational incidents [3, 6]. This research develops a new type of betweenness centrality measure to assess the transmission of infectious diseases in public transportation networks. Three concepts are introduced to study the networks where transmission of disease occurs at their edges: source-node Total Betweenness Centrality (TBC), target-node TBC, and network Encounter Matrix (EM). For every node of the network, the shortest paths tree or sub-network consists of all shortest path(s) from that node to all other network nodes. The source node TBC is defined as the sum of the BC of the edges that belong to this sub-graph. By the same token, the target node TBC of a node could be defined based on the sub-graph shortest paths from all network nodes to that node. Based on the TBC measure, we can determine the stations of those passengers who are exposed to more encounters with other passengers and consequently are more susceptible to transmission of infectious diseases. The EM is defined between each pair of source-target nodes (stations) to measure encounter opportunities in the network, which is achieved by extracting the intersection of shortest paths sub-graphs from (or to) those stations. Comparing the TBC value, the higher value of the source node TBC reveals a greater spreading influence, and the higher target node TBC shows a higher risk of infection.

Since the spread of viruses can occur at the network's nodes and edges, we develop the encounter network where every source-target node pair of the original network is a node in the encounter network. Two nodes are adjusted if and only if the corresponding shortest paths between them in the original node pairs intersect at least on one edge. The proposed method is implemented and tested on the Sioux Falls network.

2 Total Betweenness Centrality and Encounter Network

2.1 Total Betweenness Centrality

The complex networks theory has excellent applicability in describing different phenomena and has received much attention in recent years [5]. In this context, various centrality measures for quantifying the network structure have been developed and discussed [1]. The node (edge) betweenness centrality measures the intermediary of a node (edge) and the shortest path between all node pairs, thus it characterizing the importance of a node or link in flow organization in the network [1].

Total betweenness centrality (TBC) is a related concept that has been introduced in Network Science [4, 8]. In the prior work, let's denote W as the subset of a network V , the TBC of the subset $W \subseteq V$ is defined as the sum of the betweenness centrality values of its nodes, i.e., $C(W) = \sum_{i \in W} C(i)$. However, in the case of transmission of diseases, this measure may not be very useful. Human encounters of indefinite length in the same enclosed space on network edges might impose a higher risk than brief passing at nodes. An edge with a higher BC means more chances of encountering travelers from many different routes and, consequently, a higher risk of transmitting disease between people. Thus, this research modifies the traditional definition of TBC by considering the BC scores of edges. To reckon with the directionality of transportation links, we distinguish between source TBC and target TBC as defined below.

► **Definition 1.** *Source TBC*: In a directed network V , for a given source node r , the source TBC is defined as the sum of BC values of all edges of the subgraph, $\Psi^s(r)$, that consists of shortest paths from r to all other nodes in V . It can be expressed mathematically as follows.

$$TC^s(r) = \sum_{k \in \Psi^s(r)} C(k) = \sum_{k \in \Psi^s(r)} \sum_{s \neq t} \frac{\sigma_{st}(k)}{\sigma_{st}} \quad (1)$$

► **Definition 2.** *Target TBC*: In a directed network V , for a given target node r , the target TBC is defined as the sum of BC values of all edges of the subgraph, $\Psi^t(r)$, that consists of shortest paths from all other nodes to r . It can be expressed mathematically as follows.

$$TC^t(r) = \sum_{k \in \Psi^t(r)} C(k) = \sum_{k \in \Psi^t(r)} \sum_{s \neq t} \frac{\sigma_{st}(k)}{\sigma_{st}} \quad (2)$$

An illustration is provided in Figure 1 using the Sioux Falls network. The directed network consists of 24 nodes and 76 edges. In Figure 1(a), the sub-graph $\Psi^s(r)$ is in bold red color. In Figure 1-b, the sub-graph $\Psi^t(r)$ is shown in bold blue color. The two sub-graphs display all the edges where travelers from Source Node 1 might encounter travelers going to Target Node 18.

The volume and duration of encounters on each network edge could be different, subject to the travel time and trip volume on each. The following equations formalize the calculation of TBC values with selected weight:

$$TC_w^s(r) = \sum_{k \in \Psi^s(r)} w_k C(k) = \sum_{k \in \Psi^s(r)} \sum_{s \neq t} w_k \frac{\sigma_{st}(k)}{\sigma_{st}} \quad (3)$$

$$TC_w^t(r) = \sum_{k \in \Psi^t(r)} w_k C(k) = \sum_{k \in \Psi^t(r)} \sum_{s \neq t} w_k \frac{\sigma_{st}(k)}{\sigma_{st}} \quad (4)$$

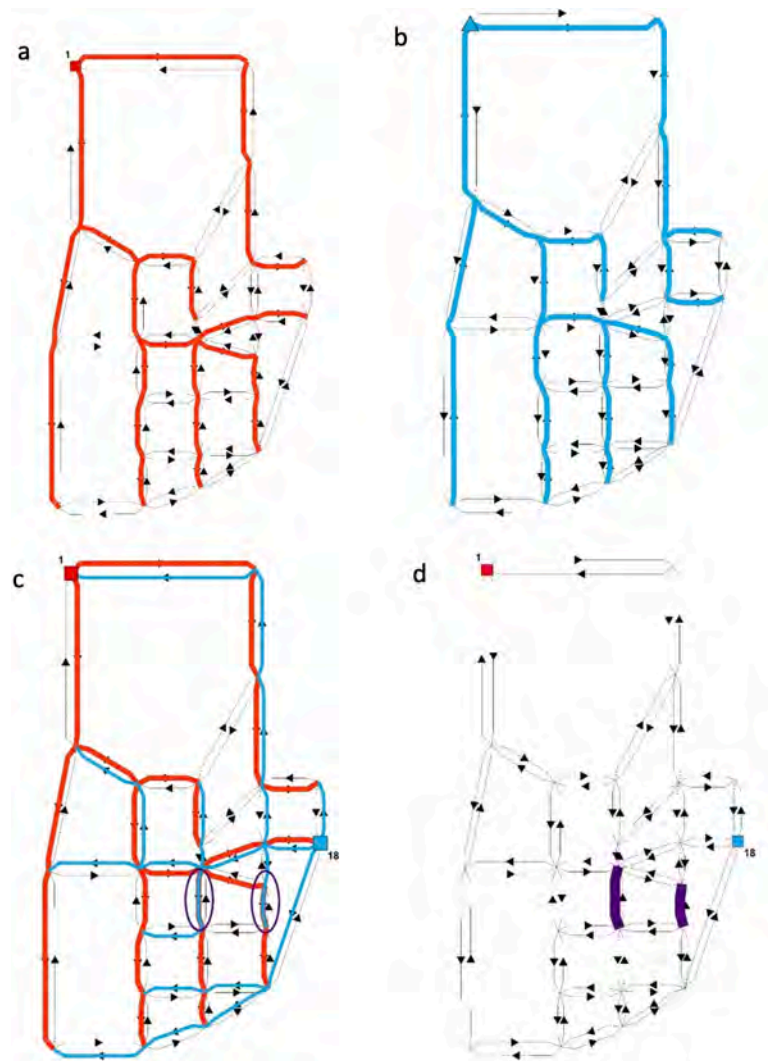
2.2 Encounter Network

► **Definition 3.** *Encounter subgraph*: the intersection sub-graph of two directed shortest-path trees, either from the source node or to a target node, is defined as the encounter subgraph.

An encounter network is constructed with the Encounter subgraph of every pair of nodes in the transit network. Consider $G_o(N_o, \epsilon_o, W_o)$, the original weighted directed graph with total N node number, where $N_o = i_1, i_2, \dots, i_N$ is the node set, ϵ_o is the edge set, and W_o is the weight set. The encounter network consists of all source-target node pairs, where each source-target pair represent one node in the encounter network. The total nodes of the encounter network would be $N \cdot (N - 1)$, and the network could be represented with $G_e(N_e, \epsilon_e, W_e)$, where $N_e = \{I_1, I_2, \dots, I_N(N - 1)\}$ is the node-set, ϵ_e and W_e are the edge and the weight sets, respectively. In the encounter network, two nodes are neighbors if and only if the shortest path(s) between the original source-target node pairs pass through at least one shared edge of the original network. Denote two arbitrary nodes in the encounter network as K and J , related to the node pairs $\{k - k'\}$ and $\{j - j'\}$, respectively. The edge between these two nodes in the encounter network is denoted as E_{KJ} , and is defined as:

$$E_{KJ} = \begin{cases} 0 & \text{if } \sigma_{kk'} \cap \sigma_{jj'} = \emptyset \\ 1 & \text{if } \sigma_{kk'} \cap \sigma_{jj'} \neq \emptyset \end{cases} \quad (5)$$

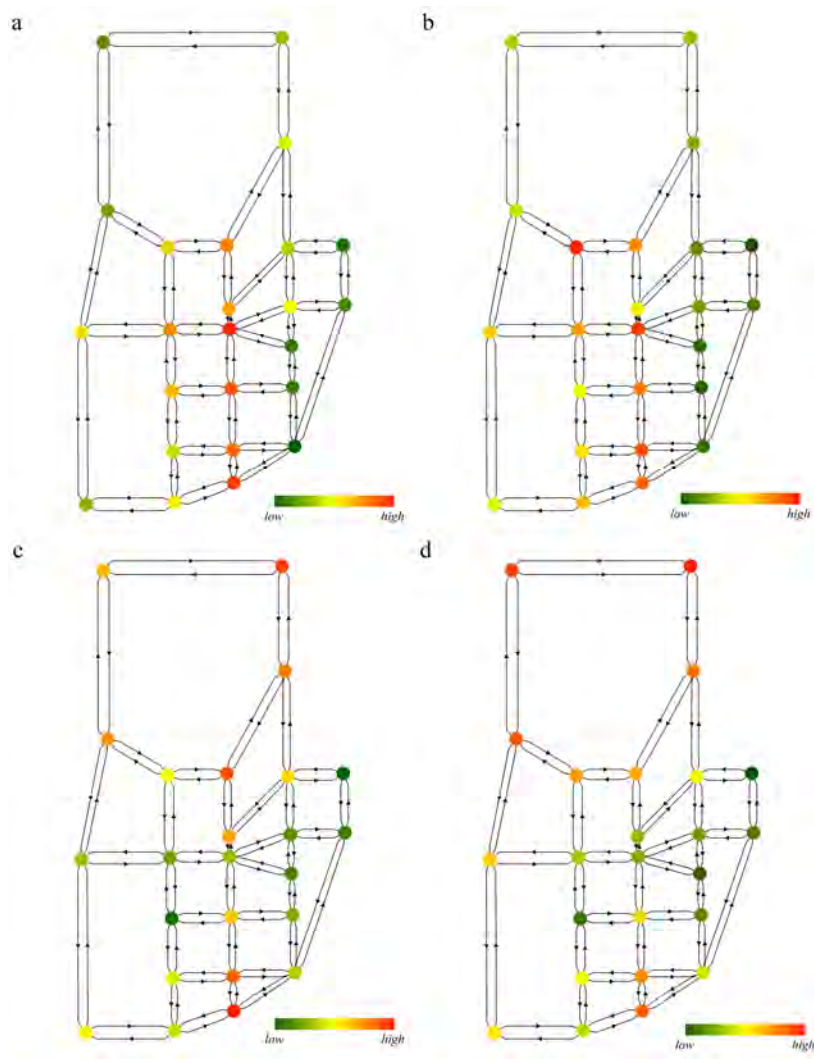
The encounter network can also be either unweighted or weighted. Depending on whether the original network is unweighted or weighted. The weight set of the network can take two different values.



■ **Figure 1** Example graph. a). Shortest path tree from the source node 1 to all other nodes; b) shortest path tree from all nodes to the target node 1; c) two shortest path trees from two source nodes 1 and 18; d) intersection sub-graph of the two shortest paths trees.

3 Preliminary results of model testing and validation

Encounter network and the TBC measures provide a structure to study epidemic spreading in the networks where the transmission occurs on network edges. To test the feasibility, we adopt the SIR epidemic model to mimic the network's spreading process. Since all contacts are not equally facilitating contagion [12], to make the simulation more compatible with the natural spreading process, the network's weight structure (e.g. passenger volume, etc) is considered to investigate the network nodes' spreading capability. Previous studies have assumed different transmission forms such as linear [9] or nonlinear [13] to model the infection rate based on the edges' weight. Here, a linear transmission rate is applied to calculate the probability of infection of a susceptible node by an infected node. The work is ongoing and reported here are the calculated source and target TBC values for each node, as shown in Figure 2. The figure shows the source and target TBCs of the Sioux Falls network for both unweighted and weighted based on the edges' travel time.



■ **Figure 2** TBC results for Sioux Fall network. a) source node TBC, b) source node weighted TBC, c) target node TBC, d) target node weighted TBC.

4 Conclusions and Discussions

Recent infectious disease outbreaks have demonstrated the vulnerability of human communities. Human encounters are a primary medium for the spreading of contagious diseases. In urban areas, public transportation systems can transfer infected people in the network and provide a conducive environment for transmitting disease through direct (Person-to-person contact) and indirect (airborne transmission or touching contaminated objects) contact between passengers. This research extended from the betweenness centrality measures and defined new TBC measures and a new concept of encounter network. They can be used to represent and model encounter opportunities on a network.

The work presents a novel method to identify more influential nodes/edges that are more likely to be the source of spreading and higher-risk nodes/edges that are more likely to receive infection. Moreover, beyond the application to infectious diseases, encounter network TBC measures can be used to study the communication characteristics of transportation networks and other complex networks like social networks.

References

- 1 Marc Barthélemy. Spatial networks. *Physics reports*, 499(1-3):1–101, 2011.
- 2 Paola Carolina Bueno, Juan Gomez, Jonathan R Peters, and Jose Manuel Vassallo. Understanding the effects of transit benefits on employees' travel behavior: Evidence from the new york-new jersey region. *Transportation Research Part A: Policy and Practice*, 99:1–13, 2017.
- 3 Robin de Regt, Christian von Ferber, Yuriy Holovatch, and Mykola Lebovka. Public transportation in great britain viewed as a complex network. *Transportmetrica A: Transport Science*, 15(2):722–748, 2019.
- 4 Martin G Everett and Stephen P Borgatti. The centrality of groups and classes. *The Journal of mathematical sociology*, 23(3):181–201, 1999.
- 5 Flavio Iannelli, Andreas Koher, Dirk Brockmann, Philipp Hövel, and Igor M Sokolov. Effective distances for epidemics spreading on complex networks. *Physical Review E*, 95(1):012313, 2017.
- 6 Qing-Chang Lu. Modeling network resilience of rail transit under operational incidents. *Transportation Research Part A: Policy and Practice*, 117:227–237, 2018.
- 7 Baichuan Mo, Kairui Feng, Yu Shen, Clarence Tam, Daqing Li, Yafeng Yin, and Jinhua Zhao. Modeling epidemic spreading through public transit using time-varying encounter network. *Transportation Research Part C: Emerging Technologies*, 122:102893, 2021.
- 8 Ahmad K Naimzada, Silvana Stefani, and Anna Torriero. *Networks, topology and dynamics: Theory and applications to economics and social systems*, volume 613. Springer Science & Business Media, 2008.
- 9 Prapanporn Rattana, Konstantin B Blyuss, Ken TD Eames, and Istvan Z Kiss. A class of pairwise models for epidemic dynamics on weighted networks. *Bulletin of mathematical biology*, 75:466–490, 2013.
- 10 Lijun Sun, Kay W Axhausen, Der-Horng Lee, and Xianfeng Huang. Understanding metropolitan patterns of daily encounters. *Proceedings of the National Academy of Sciences*, 110(34):13774–13779, 2013.
- 11 Behnam Tahmasbi and Hossein Haghshenas. Public transport accessibility measure based on weighted door to door travel time. *Computers, Environment and Urban Systems*, 76:163–177, 2019.
- 12 Riitta Toivonen, Jussi M Kumpula, Jari Saramäki, Jukka-Pekka Onnela, János Kertész, and Kimmo Kaski. The role of edge weights in social networks: modelling structure and dynamics. In *Noise and Stochastics in Complex Systems and Finance*, volume 6601, pages 48–55. SPIE, 2007.
- 13 Wei Wang, Ming Tang, Hai-Feng Zhang, Hui Gao, Younghae Do, and Zong-Hua Liu. Epidemic spreading on complex networks with general degree and weight distributions. *Physical Review E*, 90(4):042803, 2014.
- 14 Hai Yang and Hai-Jun Huang. *Mathematical and economic theory of road pricing*. Emerald Group Publishing Limited, 2005.

Inferring the History of Spatial Diffusion Processes

Takuya Takahashi ✉ 

Department of Geography, University of Zurich, Switzerland

Geneviève Hannes ✉ 

Department of Geography, University of Zurich, Switzerland

Nico Neureiter ✉ 

Department of Geography, University of Zurich, Switzerland
NCCR Evolving Language, University of Zurich, Switzerland

Peter Ranacher ✉ 

URPP Language and Space, University of Zurich, Switzerland
Department of Geography, University of Zurich, Switzerland
NCCR Evolving Language, University of Zurich, Switzerland

Abstract

When studying the spatial diffusion of a phenomenon, we often know its geographic distribution at one or more snapshots in time, while the complete history of the diffusion process is unknown. For example, we know when and where the first Indo-European languages arrived in South America and their current distribution. However, we do not know the history of how these languages spread, displacing the indigenous languages from their original habitat. We present a Bayesian model to interpolate the history of a diffusion process between two points in time with known geographical distributions. We apply the model to recover the spread of the Indo-European languages in South America and infer a posterior distribution of possible evolutionary histories of how they expanded their areas since the time of the first invasion by Europeans. Our model is more generally applicable to infer the evolutionary history of geographic diffusion phenomena from incomplete data.

2012 ACM Subject Classification Computing methodologies

Keywords and phrases Bayesian inference, geographic diffusion, language evolution, Indo-European, colonisation of the Americas

Digital Object Identifier 10.4230/LIPIcs.GIScience.2023.71

Category Short Paper

Funding Funding supports for this work were provided by the URPP Language and Space, University of Zurich, the NCCR Evolving Language with Swiss NSF Agreement No. 51NF40_180888, and the Swiss NSF Sinergia Project No. CRSII5_183578 (Out of Asia).

Acknowledgements We thank Gereon Kaiping for valuable discussion and ideas in the early phase of the project.

1 Introduction

Following the European colonisation of the Americas during the Age of Discovery, Indo-European (IE) languages, such as Spanish, Portuguese and French, spread extensively in South America, eliminating many indigenous languages. Historical records show when and where the IE languages arrived on the continent. The current spatial distribution of languages in South America is available in modern language maps. However, little is known about the spatio-temporal diffusion of the IE languages between the time of first contact at about 1500 CE and today. How have the IE languages spread between the time of contact and today? How can we infer probable evolutionary histories of this diffusion process in the absence of relevant historical records?



© Takuya Takahashi, Geneviève Hannes, Nico Neureiter, and Peter Ranacher;
licensed under Creative Commons License CC-BY 4.0

12th International Conference on Geographic Information Science (GIScience 2023).

Editors: Roger Beecham, Jed A. Long, Dianna Smith, Qunshan Zhao, and Sarah Wise; Article No. 71; pp. 71:1–71:6
Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

In GIScience, similar questions frequently arise when studying diffusion phenomena, such as urban sprawl, deforestation, land cover change, segregation, or the spread of innovation. We know the spatial distribution at two points in time, and we would like to interpolate potential histories of how the diffusion has unfolded in between.

Various methods have been used to model the diffusion process in space. Cellular automata (CA) were applied to simulate the urban sprawl [4]. Reaction-diffusion equations were used to represent the demic-diffusion of modern humans theoretically [7]. Network models were used to describe the diffusion of human culture [6] and dialects [5]. Ising models were used to describe the diffusion of linguistic features in the UK [2, 3]. While these models can simulate the diffusion process from a given initial spatial distribution, they cannot infer the evolutionary history between two known points in time.

In this paper, we present a novel Bayesian model to interpolate potential histories of a spatial diffusion process, capturing the uncertainty of the process. The model reveals the distribution of the most likely evolutionary histories of a diffusion process, given the spatial distribution of the process at two points in time.

We applied the model in a case study to interpolate the spatial diffusion of the invasive IE languages in the Americas between the time of contact and today, capturing the uncertainty of the process. The model gives the posterior probability that an IE language occupied a given location in South America at a given time.

2 Methods

2.1 Model assumptions

We represent space as a network of n discrete nodes P_1, \dots, P_n , each assigned to one of K possible states. In the case study, P_i is a cell in a regular spatial grid over South America. The cell has two states: 1 means an IE language occupies the cell, and 0 means an indigenous language occupies the cell. We denote the geographical distribution of states at time t with the vector

$$\mathbf{M}(t) = \begin{pmatrix} x_1(t) \\ \vdots \\ x_n(t) \end{pmatrix},$$

where $x_i(t)$ is the state of node P_i at time t . In the case study, $\mathbf{M}(t)$ is the geographical distribution of the IE languages in South America at a specific time in history. We model the spatial diffusion as a Markov process, where $\mathbf{M}(t)$ only depends on $\mathbf{M}(t-1)$ and is independent of earlier time steps. At each time t , every node copies the state from its neighbours at time $t-1$. The transmission rate a_{ij} , with $0 \leq a_{ij} \leq 1$ and $\sum_{j=1}^n a_{ij} = 1$, gives the probability that node P_i copies the state from P_j . The transmission rate is a constant and must be defined before the analysis. In the case study, each cell can copy the state from its eight neighbours and itself with equal probability:

$$a_{ij} = \begin{cases} \frac{1}{9} & \text{if } P_i = P_j \text{ or } P_i \text{ and } P_j \text{ are neighbours} \\ 0 & \text{otherwise} \end{cases}.$$

2.2 Bayesian inference

We use Bayesian inference to estimate the spatial diffusion $\mathbf{M}(0), \dots, \mathbf{M}(T)$, i.e. the history of the geographic distribution of states. The spatial diffusion follows a Markov process and has the probability

$$P(\mathbf{M}(0), \dots, \mathbf{M}(T)) = P(\mathbf{M}(0)) \prod_{t=1}^T P(\mathbf{M}(t) | \mathbf{M}(t-1)).$$

We know the geographic distribution at two points in time, the initial distribution at $t = 0$ and the final distribution at $t = T$. The history between initial and final distribution, $\mathbf{M}(1), \dots, \mathbf{M}(T - 1)$, has posterior probability

$$\begin{aligned} P(\mathbf{M}(1), \dots, \mathbf{M}(T - 1) \mid \mathbf{M}(0), \mathbf{M}(T)) &= \frac{P(\mathbf{M}(0), \dots, \mathbf{M}(T))}{P(\mathbf{M}(0), \mathbf{M}(T))} \\ &= \frac{P(\mathbf{M}(0))}{P(\mathbf{M}(0), \mathbf{M}(T))} \prod_{t=1}^T P(\mathbf{M}(t) \mid \mathbf{M}(t - 1)) \\ &\propto \prod_{t=1}^T \prod_{i=1}^n P(x_i(t) \mid \mathbf{M}(t - 1)). \end{aligned} \quad (1)$$

2.3 Markov chain Monte Carlo (MCMC)

We can use the Metropolis-Hasting (M-H) algorithm to draw samples from the posterior distribution in Equation 1, repeating the following steps:

1. Randomly choose one timestep t and one node P_i .
2. If $x_i(t) = k$, propose k' as a candidate state with the proposal distribution

$$q(k' \mid k) = \begin{cases} \frac{1}{K-1} & \text{if } k' \neq k \\ 0 & \text{otherwise} \end{cases}.$$

3. Compute the acceptance ratio

$$r = \frac{P(x_i(t) = k' \mid \mathbf{M}(t - 1))}{P(x_i(t) = k \mid \mathbf{M}(t - 1))} \prod_{j \in N(i)} \frac{P(x_j(t + 1) \mid x_i(t) = k', \cap_{(1 \leq l \leq n, l \neq i)} x_l(t))}{P(x_j(t + 1) \mid x_i(t) = k, \cap_{(1 \leq l \leq n, l \neq i)} x_l(t))}, \quad (2)$$

where $N(i)$ is the set of neighbours of P_i , or formally the set $\{j \mid 1 \leq j \leq n, a_{ji} > 0\}$. The conditional probabilities in expression 2 are computed with the transmission rates a_{ij} .

4. Accept the proposal with probability $\min(r, 1)$.

Letting m denote the average node degree of the network, one iteration of the MH-algorithm runs in $O(m)$ time.

3 Case study

In this section, we apply our model to explore the diffusion of the IE languages in South America, and interpolate probable evolutionary histories of how they have expanded their geographical area.

3.1 Network and diffusion model

We segmented the landmass of South America into a regular grid, each grid cell representing a node in the network. The Moore neighbourhood gives the transmission rate between cells: each cell may copy the state from its eight neighbours or itself with equal probability. Cells can take two states:

$k = 0$... an indigenous language occupies the cell

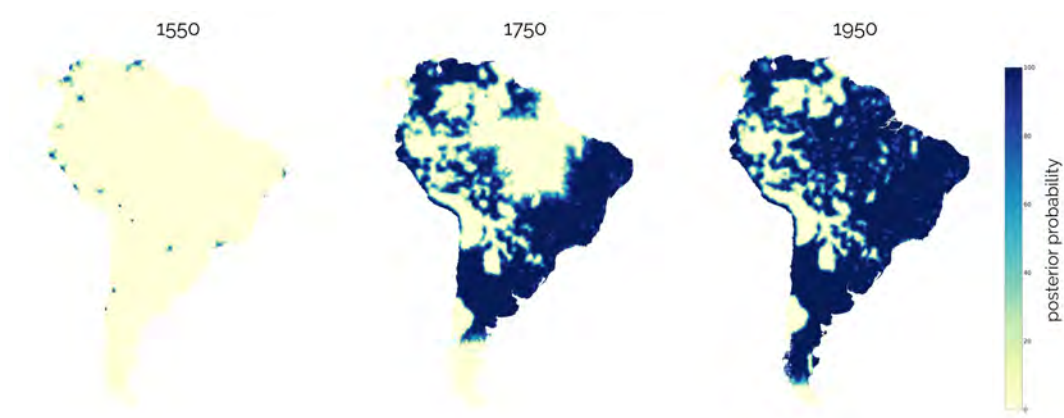
$k = 1$... an IE language occupies the cell

3.2 Data

The data comprise the geographical distributions of languages in 1510, the time of the first invasion, and 1990, the modern geographical distribution[1]. We included additional European arrivals between 1510 and 1990 from the literature. For example, the Spanish arrived in Santa Marta, modern-day Colombia, in 1525, and we fixed the state of the corresponding grid cells to 1 in the spatial distribution for this year.

3.3 Results

Figure 1 shows the posterior probability of the IE languages reaching each grid cell by 1550, 1750, and 1950. The IE languages gradually spread inland from the initial points of arrival at the coast. Figure 2 shows the posterior distribution of the arrival of the IE languages to selected cities along the Amazon basin.



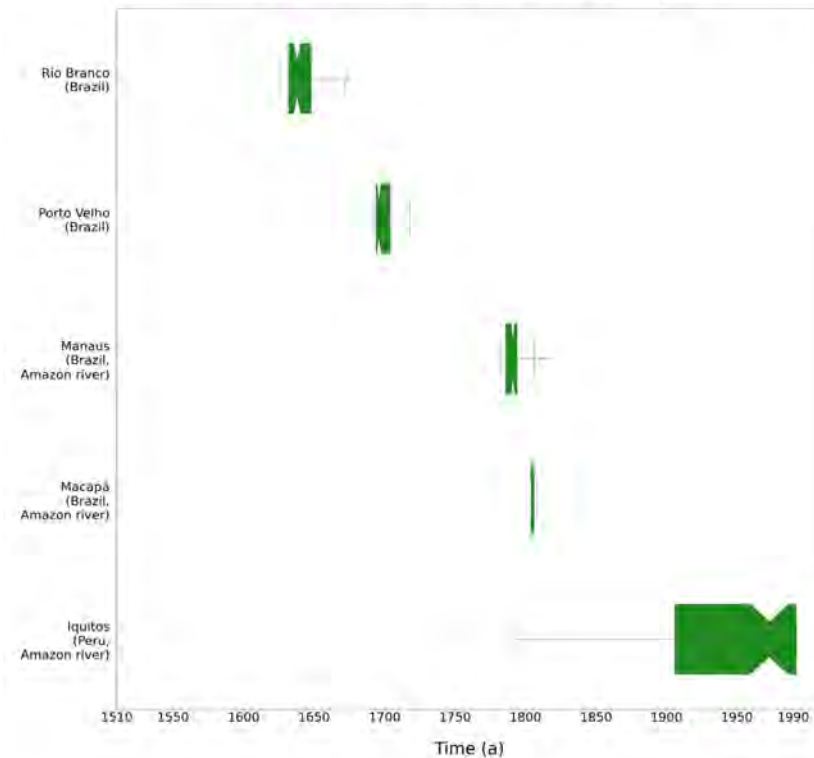
■ **Figure 1** Posterior distribution of IE languages reaching each grid cell by 1550, 1750, and 1950.

4 Discussion

In this paper, we presented a Bayesian model to interpolate the evolutionary history of a spatial diffusion process between two points in time with known geographic distributions. In a case study, the model showed likely scenarios of how the invasive Indo-European languages drove the indigenous languages of South America out of their original habitat.

In contrast to the conventional CA models, the model is fully Bayesian and returns a posterior distribution of possible evolutionary histories instead of just a single best history. In the case study, the model revealed the posterior probability of the IE languages reaching locations in South America between 1510 and 1990. Moreover, one can easily add prior information to Bayesian models and estimate the posterior distribution of potential evolutionary histories considering all available knowledge in a principled way. In the case study, for example, we added the locations of additional European entries to South America between the two known times in history.

In our model, the transmission rate reflects the influence of Geography on spatial diffusion. Since Bayesian models return a full posterior distribution, we can compare models with different transmission rates, e.g. using the Bayes factor, and evaluate the effect of geographic hypotheses on the diffusion process. For example, geographical barriers such as mountains and rivers might hinder the diffusion of languages, blocking the displacement of human groups. We could model this influence with lower transmission rates in mountainous terrain.



■ **Figure 2** Posterior distribution of the arrival of IE to selected cities.

Another possible extension includes mutation events, where a node may acquire a state not shared by any of its neighbours with a non-zero probability. Modelling the mutation event will enable the inference of an unrecorded arrival of an IE language not included in the data. Comparing two models with and without the mutation event could show whether today's geographical distribution has been formed by continuous diffusion or discontinuous state change. Since the geographical distribution at a given time still only depends on that at the previous time, including the mutation event does not violate the assumptions of a Markov process.

5 Conclusion

We present a method to infer potential histories of a spatial diffusion process between two points in time with known spatial distributions. We applied the method to infer the history of the IE languages spreading and displacing the indigenous languages in South America. Our method is more broadly applicable to infer the evolutionary history of geographic diffusion phenomena from incomplete data, frequently occurring in GIScience.


References

- 1 Ronald E Asher and Christopher Moseley. *Atlas of the world's languages*. Routledge, 2018.
- 2 James BurrIDGE. Spatial evolution of human dialects. *Phys. Rev. X*, 7:031008, July 2017. doi:10.1103/PhysRevX.7.031008.

71:6 Inferring the History of Spatial Diffusion Processes

- 3 James Burridge and Tamsin Blaxter. Using spatial patterns of english folk speech to infer the universality class of linguistic copying. *Phys. Rev. Res.*, 2:043053, October 2020. doi:10.1103/PhysRevResearch.2.043053.
- 4 Lingling Sang, Chao Zhang, Jianyu Yang, Dehai Zhu, and Wenju Yun. Simulation of land use spatial pattern of towns and villages based on ca-markov model. *Mathematical and Computer Modelling*, 54(3):938–943, 2011. Mathematical and Computer Modeling in agriculture (CCTA 2010). doi:10.1016/j.mcm.2010.11.019.
- 5 Takuya Takahashi and Yasuo Ihara. Quantifying the spatial pattern of dialect words spreading from a central population. *Journal of The Royal Society Interface*, 17(168):20200335, 2020. doi:10.1098/rsif.2020.0335.
- 6 Takuya Takahashi and Yasuo Ihara. Application of a markovian ancestral model to the temporal and spatial dynamics of cultural evolution on a population network. *Theoretical Population Biology*, 143:14–29, 2022. doi:10.1016/j.tpb.2021.10.003.
- 7 Joe Yuichiro Wakano, William Gilpin, Seiji Kadowaki, Marcus W. Feldman, and Kenichi Aoki. Ecocultural range-expansion scenarios for the replacement or assimilation of neanderthals by modern humans. *Theoretical Population Biology*, 119:3–14, 2018. doi:10.1016/j.tpb.2017.09.004.

Modelling Affordances as Emergent Phenomena

Sabine Timpf¹  

Geoinformatics Group, University of Augsburg, Germany

Franziska Klügl  

AASS/NT, Örebro University, Sweden

Abstract

Affordances are an important basis for many human-environment interactions such as navigation or geo-design. In this short paper we present an approach to modelling affordances based on treating affordances as emergent phenomena in an agent-based simulation. We use the notion of an affordance schema to represent the setting in which the emergence of an affordance is made possible. We use a case study to show that (unexpected) affordances emerge during the course of the simulation. While the general approach is promising and may be used for other emergent phenomena such as landmarks, we also acknowledge and discuss the problems incurred during the modelling process. The paper closes with a reflection and some ideas for future work.

2012 ACM Subject Classification Computing methodologies → Modeling methodologies; Computing methodologies → Spatial and physical reasoning

Keywords and phrases agent-based modelling, cognitive engineering, spatial cognition, theory of modelling

Digital Object Identifier 10.4230/LIPIcs.GIScience.2023.72

Category Short Paper

Supplementary Material *Software*: <https://github.com/sabinetimpf/emergentAffordance>

Funding *Sabine Timpf*: Funding from the Erasmus+ program is gratefully acknowledged.

Acknowledgements The constructive comments from two reviewers are gratefully acknowledged.

1 Understanding human-environment interaction: the modelling challenge

In this short paper we present our recent work on how to model affordances, which are important for many human-environment interactions such as for example navigation or geo-design. Understanding how humans interact with their environment is part of understanding decision-making. Affordances, according to Gibson [3], are what the environment offers the individual in terms of interaction. Seen from the individual's perspective, affordances represent potential actions tied to specific objects, subjects or groups of objects, but also tied to the current status, knowledge or beliefs of the individual.

Modelling affordances is not a new endeavour and a proper overview does not fit into this paper. Extensions such as [10] introduce cognition into the theory of affordances. However, all implemented approaches so far (see for example [14], [17] or [8]) treat affordances as properties or functions of an object or as properties of the individual-object relationship as a whole. This solution does not satisfy the emergent nature of affordances as described by Gibson.

Sahin et al. [13] define affordances from an agent's perspective within the context of robot control. Their approach is different in that it acknowledges the dynamic nature of the affordance and treating it as a relation between equivalence classes. In robotics,

¹ corresponding author



© Sabine Timpf and Franziska Klügl;
licensed under Creative Commons License CC-BY 4.0

12th International Conference on Geographic Information Science (GIScience 2023).

Editors: Roger Beecham, Jed A. Long, Dianna Smith, Qunshan Zhao, and Sarah Wise; Article No. 72; pp. 72:1–72:6

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

computational models of affordances have gained a new impetus as evidenced by several overviews in recent years ([19], [1]) and a recent special issue ([11]). However, these approaches do not (yet) provide high-level action information, remaining at the level of interpreting sensory-motor sensor input, which is very different from our focus on understanding and modelling affordances in human-environment interaction.

According to Gibson [3] affordances are perceived immediately without any reasoning, i.e. they emerge from the dynamic relationship between individual and object. While Gibson hypothesised that this perception is inborn, other ecological psychologists such as Neisser [9] argued that the perception of action potentials is a result of a cyclical learning process. While this discussion needs to be solved (preferably by psychologists), our concern is with the modelling of the, let's call it, mechanism of how affordances are supposed to work, assuming that affordances are emergent phenomena as posited by Gibson.

The moniker of emergence originally stems from systems theory, where it describes an observable phenomenon that was not originally visible or predictable by observing the different parts that constitute a system. In our case the system consists of a human and an environmental object (object collection) as well as their interaction. The interaction is the observable result of realising the action potential of the affordance. There may be several affordances providing distinct action potentials in any given agent-object pairing.

From a computational point of view the question arises how an emergent phenomenon may be modelled at all. By its nature a phenomenon emerges during the model run and should not be explicitly represented in our model. Can we then model a system in which we increase the probability of an emergent phenomenon occurring? We approach this question by changing from an analytic (property-oriented) paradigm to an agent-based constructive paradigm.

2 Changing the perspective: an agent-based approach

Agent-based modelling has been shown to be able to capture emergent phenomena resulting from the interactions of individual entities [2]. Emergent phenomena in agent-based models are patterns, structures and behaviours that were not explicitly implemented in the model, but arise through agents' interaction. An agent-based model consists of dynamically interacting agents that use rules for their own behaviour [7]. Such models are commonly used to analyse complex systems that are characterised by a large variety of components that interact with each other. In the case of modelling for emergent affordances, we need to determine the rules agents follow that require some interaction with an entity, the constraints under which an interaction may take place and the entities that may represent interaction partners. This kind of interaction follows a pattern that is akin to a schema in cognitive science [4].

2.1 Using schemata to model affordances

Neisser [9] states that humans use schemata to make sense of their surroundings and to minimise the facts they need to memorise. Rumelhart calls schemata the “building blocks of cognition” [12]. Consider that you know the schema underlying the concept of a “bridge”, then there is no need to memorise every bridge that you encounter. You will have learned that a bridge is an instance of the link schema that allows you to connect and move between two areas using a direct path. Schemata are recurring structures we learn that help us to establish patterns of understanding and reasoning.

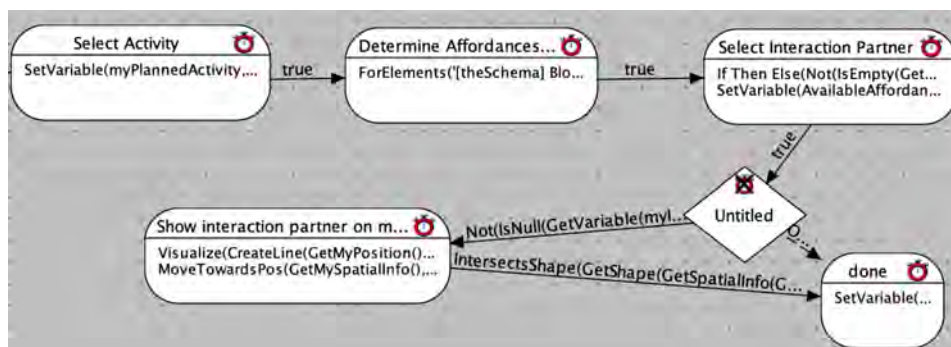
We will show that the implementation of an affordance schema provides an answer to the modelling challenge, i.e. the emergent nature of affordances. In this implementation the affordance schema serves as a kind of template for generating an affordance at simulation

run-time. While the notion of an affordance schema originally served as a means to make interactions visible and tractable [5], we can show in this paper that it also satisfies the question of how to model emergent phenomena. The notion of a schema allows for the required flexibility in matching the needs of the agent with the required properties of the environment, thus not only allowing a single affordance to emerge, but producing a collection of potential actions, including those that are unexpected or may be wrong (false affordances). As Withagen et.al. [18] discuss, there must be a way of capturing that affordances also invite behaviour not merely provide potential actions.

In our implementation, agents possess a list of affordance schemata for each of the activities they may carry out. We define an affordance schema as a 3-tuple $\langle EType, condition, fpriority \rangle$ composed of

- an Entity type EType,
- a condition that expresses constraints under which the affordance can be generated, and
- a number called fpriority that assigns a priority or preference to the combination of agent and potential interaction partners.

During run-time an affordance $\langle a, e, act, p \rangle$ is generated, where a stands for the agent, e stands for environmental object, act describes the activity the agent intends to perform and p stands for preference, allowing a means of differentiating between otherwise equivalent affordances.



■ **Figure 1** Activity graph of affordance-enabled agent.

Figure 1 shows the activity graph of the agent where, after selecting an activity, first the affordances are determined and then the interaction partner is selected. A detailed account of the implementation may be found in [6] and on github².

2.2 Case study: a visit to the park

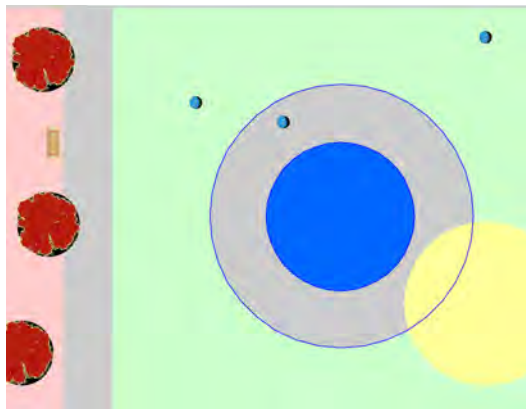
We are applying the approach discussed above to a case study of visiting a public park. We focus on the act of walking into a park and needing to find a place to perform a specific activity there ([15], [16]). As an example we take the specific activity of sitting to take a break with the option of drinking some water. The question arises which objects in the park offer the affordance for sitting as well as drinking some water. As shown in [6] it is necessary to break down the conditions for the activity in terms of attributes of the environmental object tempered by knowledge about the agent. For example, in order to be able to sit, there

² <https://github.com/sabinetimpf/emergentAffordances>

72:4 Modelling Affordances as Emergent Phenomena

must be an object or configuration there that affords sitting, which means a relatively flat, stable surface with a certain minimal size. There should be a differentiation into required and optional properties.

In contrast to an implementation based purely on an object's properties, in our implementation the properties are expressed as a function of the agent's properties, i.e. the 'certain size' is expressed in terms of an agent's size, the 'stable surface' may take the agent's weight as parameter, and the 'relatively flat surface' may take preferences of the agent into account. Please note that this customisation is only possible because of the agent-based approach that allows tying agent and interaction object together at run-time. This allows for the affordance to truly emerge as an individual trait between agent and object.



■ **Figure 2** Situation of park scenario at run-time.

Figure 2 shows a detail of the simulation situation, where an agent has chosen to sit on a relatively low concrete band surrounding a water feature. This interaction partner is unusual but perfectly fine for the purpose of sitting and drinking some water. Of course, we did not specify that the water in the water feature might not be safe to drink for humans.

While the current results of the implementation are encouraging, we must note that it is quite time-consuming to put into constraints all required and optional properties of agent and objects within an activity context for physical affordances; And as the example shows, it is easy to forget specific aspects. However, we believe that this example shows the flexibility of using affordance schemata to model human-environment interactions that also allows the emergence of affordances in the original sense of Gibson, as we interpret it.

3 Reflections and future work

In this research-in-progress we have used affordance schemata as patterns that generate affordances during an agent-based simulation. This version of modelling affordances seems to be closer to the original idea of Gibson, who saw affordances as emerging phenomena. We have implemented this approach, thus showing proof of concept. We are currently working on a more detailed and extensive implementation of the park use scenario. However, there should be a better way of defining the constraints and ensuring their completeness.

One endeavour for the future is to formally define activities and their actions as well as the needed properties for an activity to be carried out successfully. While we have extensive observations of park behaviour to help us with the formalisation, this might not be true for other spatial behaviours. Our approach using schemata is promising also for modelling other emergent phenomena such as landmarks or resilient objects, which is an avenue we would like to explore in the future.




References

- 1 Paola Ardón, Eric Pairet, Katrin S Lohan, Subramanian Ramamoorthy, and Ronald Petrick. Building affordance relations for robotic agents—a review. *arXiv preprint arXiv:2105.06706*, 2021.
- 2 Eric Bonabeau. Agent-based modeling: Methods and techniques for simulating human systems. *PNAS*, 99(3):7280–7287, 2002.
- 3 James J. Gibson. *The Ecological Approach to Visual Perception*. Houghton Mifflin, 1979.
- 4 Mark Johnson. *The body in the mind: The bodily basis of meaning, imagination, and reason*. University of Chicago Press, Chicago, IL, US, 1987.
- 5 Franziska Klügl and Sabine Timpf. Approaching interactions in agent-based modelling with an affordance perspective. In Gita Sukthankar and Juan A. Rodríguez-Aguilar, editors, *Autonomous Agents and Multiagent Systems - AAMAS 2017 Workshops, Best Papers, São Paulo, Brazil, May 8-12, 2017, Revised Selected Papers*, volume 10642 of *Lecture Notes in Computer Science*, pages 222–238. Springer, 2017. doi:10.1007/978-3-319-71682-4_14.
- 6 Franziska Klügl-Frohnmeier and Sabine Timpf. Towards more explicit interaction modelling in agent-based simulation using affordance schemata. In *Deutsche Jahrestagung für Künstliche Intelligenz*, 2021.
- 7 C. M. Macal and M. J. North. Tutorial on agent-based modelling and simulation. *Journal of Simulation*, 4:151–162, 2010.
- 8 Raubal Martin. Human wayfinding in unfamiliar buildings: a simulation with a cognizing agent. *Cognitive Processing*, 2, January 2001.
- 9 Ulric Neisser. *Cognition and Reality: Principles and Implications of Cognitive Psychology*. Books in psychology. W. H. Freeman, 1976.
- 10 Martin Raubal. Ontology and epistemology for agent-based wayfinding simulation. *Int. Journal of Geographical Information Science*, 15:653–665, 2001.
- 11 Erwan Renaudo, Philipp Zech, Raja Chatila, and Mehdi Khamassi. Editorial: Computational models of affordance for robotics. *Frontiers in Neurorobotics*, 16, 2022. doi:10.3389/fnbot.2022.1045355.
- 12 David E. Rumelhart. Schemata: The building blocks of cognition. In R. J. Spiro, B. C. Bruce, and W. F. Brewer, editors, *Theoretical Issues in Reading Comprehension*, pages 33–58. Erlbaum, Hillsdale, NJ, 1980.
- 13 Erol Şahin, Maya Çakmak, Mehmet R. Doğar, Emre Uğur, and Göktürk Üçoluk. To afford or not to afford: A new formalism of affordances towards affordance-based robot control. *Adaptive Behavior*, 15(4):447–472, 2007.
- 14 Thomas A. Stoffregen. Affordances as properties of the animal environment system. *Ecological Psychology*, 15(2):115–134, 2003.
- 15 Sabine Timpf. Appropriating places in public spaces: a multi-agent simulation. In Franziska Klügl, Sabine Timpf, and Ute Schmid, editors, *Agent-based simulation: from cognitive modelling to engineering practice; workshop at the 31th German Conference on Artificial Intelligence*, 2008. URL: <http://ki.informatik.uni-wuerzburg.de/events/cog.abs/WS1-KI08-Proceedings.pdf>.
- 16 Sabine Timpf and Marie-Rose Degg. Agent-based modelling of people’s behaviour in public parks. In Jason Thompson, Minh Le Kieu, and Koen van Dam, editors, *ABMUS2022: The 6th International Workshop on Agent-Based Modelling of Urban Systems*, 2022. doi:10.6084/m9.figshare.19733800.
- 17 Michael T. Turvey. Affordances and prospective control: An outline of the ontology. *Ecological Psychology*, 4(3):173–187, 1992. doi:10.1207/s15326969eco0403_3.
- 18 Rob Withagen, Harjo De Poel, Duarte Araujo, and Gert-Jan Pepping. Affordances can invite behavior: Reconsidering the relationship between affordances and agency. *New Ideas in Psychology*, 30:250–258, May 2012. doi:10.1016/j.newideapsych.2011.12.003.

72:6 Modelling Affordances as Emergent Phenomena

- 19 Philipp Zech, Simon Haller, Safoura Rezapour Lakani, Barry Ridge, Emre Ugur, and Justus Piater. Computational models of affordance in robotics: a taxonomy and systematic classification. *Adaptive Behavior*, 25(5):235–271, 2017. doi:10.1177/1059712317726357.

The FogDetector: A User Survey to Measure Disorientation in Pan-Scalar Maps

Guillaume Touya   

LASTIG, Univ Gustave Eiffel, IGN-ENSG, F-77420 Champs-sur-Marne, France

Justin Berli 

LASTIG, Univ Gustave Eiffel, IGN-ENSG, F-77420 Champs-sur-Marne, France

Abstract

When we navigate into interactive multi-scale maps that we call pan-scalar maps, it is usual to feel disoriented. This is partly due to the fact that map views do not always contain visual cues of the location of the past map views of the navigation. This paper presents an online study that seeks to understand and measure this disorientation occurring when zooming in or out of a pan-scalar map. An online study was designed and more than 150 participants finished the survey. The study shows a very small difference between the time to succeed in the memorising task after a zoom and a pan, but the difference is more significant when we compare zooming in with a large scale gap to panning. The study also shows that disorientation is not similar when zooming in and zooming out.

2012 ACM Subject Classification Applied computing → Cartography

Keywords and phrases disorientation, zoom, pan, multi-scale map, desert fog, user survey

Digital Object Identifier 10.4230/LIPIcs.GIScience.2023.73

Category Short Paper

Supplementary Material *Software*: <https://doi.org/10.5281/zenodo.7561925>

Funding *Guillaume Touya*: this project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 101003012 - LostInZoom).

Justin Berli: this project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 101003012 - LostInZoom).

Acknowledgements The authors want to thank all the anonymous participants of the FogDetector survey

1 Introduction

When we use pan-scalar maps, i.e. interactive, zoomable, multi-scale slippery maps [5], it is usual to experience disorientation when zooming in and out. As the use of these pan-scalar maps is quite recent, we do not really know much about this disorientation feeling. Between two zoom levels, particularly when they are not consecutive, the style and content can change drastically, which does not completely remove the visual cues but reduces their number: zooming from one scale to the other might thus cause disorientation [13].

The consequences of pan-scalar disorientation cannot completely be compared to geographic disorientation. People do not stop using pan-scalar maps because of this disorientation. But it can force us to zoom out (or zoom in if we were zooming out), or at least cause a delay in our use of the map. Disorientation makes the pan-scalar map exploration a more tedious task. Though the need for more research on pan-and-zoom interactions with maps was identified almost twenty years ago [6], disorientation is still significant in current pan-scalar maps. Our long-term goal is to design pan-scalar maps where interactions are smooth or



© Guillaume Touya and Justin Berli;
licensed under Creative Commons License CC-BY 4.0

12th International Conference on Geographic Information Science (GIScience 2023).

Editors: Roger Beecham, Jed A. Long, Dianna Smith, Qunshan Zhao, and Sarah Wise; Article No. 73; pp. 73:1–73:6

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

■ **Table 1** Distribution of the ages of the 160 remaining participants after the cleaning step.

	18-24	25-34	35-44	45-59	60+	no answer
nb of participants	12	41	39	51	16	1

fluid [3]. But before designing better pan-scalar maps, we argue that it is necessary to better understand disorientation, to know when and how much a map reader can be disoriented. Montello defines geographic disorientation as a phenomenon occurring “when people are aware they are not certain about where they are and/or where they need to go to get to their destination.” [9]. In this definition, we can see two components of disorientation, the objective uncertainty about where we are, and the subjective awareness of this uncertainty [4]. When it comes to the virtual disorientation occurring during the exploration of a pan-scalar map, it can be modelled as a reconciliation problem between the visual cues in the current map view and the mental map of the user [13]. There are different forms of a failed reconciliation, i.e. disorientation, and different causes [2]. In this paper, we present a user survey that seeks to measure disorientation. The desert fog effect identified in human-computer interaction [7] is one of the possible causes, hence the name of the presented survey. The desert fog is the disorientation occurring in multi-scale interactive environments when the current view does not contain visual cues referring to the past views. From a cognition perspective, this disorientation could be caused by change blindness[14] as the display changes after a zoom, or inattentive blindness [11], the map readers cannot focus their vision on all the details in the map. More generally, this disorientation can be related to limits of our visual working memory [10]. The paper is structured as follows. The survey is presented in Section 2. Section 3 presents and discusses the survey results.

2 The FogDetector survey

2.1 Hypotheses

There are many interactions possible with an interactive pan-scalar map [12], but we are only interested here in the two main displacement interactions available in such maps: pan and zoom. When you switch the layers of the map or change the style of the base map, disorientation can also occur but this case is beyond the scope of this study. Based on our understanding of the disorientation phenomenon, we make the following hypotheses:

- it takes longer to know where you are after a zoom, than after a pan (H_1).
- disorientation occurs differently when zooming in and zooming out (H_2).
- there is more disorientation when the scale difference is bigger (H_3).
- the style and generalisation have an effect on the intensity of the desert fog (H_4).

2.2 Participants and Apparatus

We tried to select a purposeful sampling for the participants of the survey, i.e. a sampling that represents the envisioned end-users. As anyone can be a user of a pan-scalar map, we wanted users with very diverse ages and experiences with interactive cartography. 160 participants were recruited online. There is a fairly good distribution of ages which confirms that it is a purposeful sample (Table 1). However, the gender distribution is skewed with 98 men, 53 women and 9 who preferred not to answer.

Table 2 shows the declared usage of pan-scalar maps by the participants of the survey. The distribution is clearly skewed towards the regular use of such maps. Though we do not have data on the use of such applications by the general public, our sample seems to use

■ **Table 2** Distribution of the declared usage of pan-scalar maps by the 160 remaining participants after the cleaning step.

	every day	once a week	once a month	almost never	no answer
nb of participants	82	55	16	4	3



■ **Figure 1** Three of the chosen targets, a building at zoom level 17 on the left, a crossroad at zoom level 12 in the middle, and a point of interest (a fountain in a square) at zoom level 17 on the right.

maps more regularly. Both to deal with potential COVID-related limitations and to access our purposeful sample, we opted for a fully online survey, based on a web application. The code of the application is openly accessible on Github¹. The data are collected anonymously to follow the guidelines of GDPR legislation.

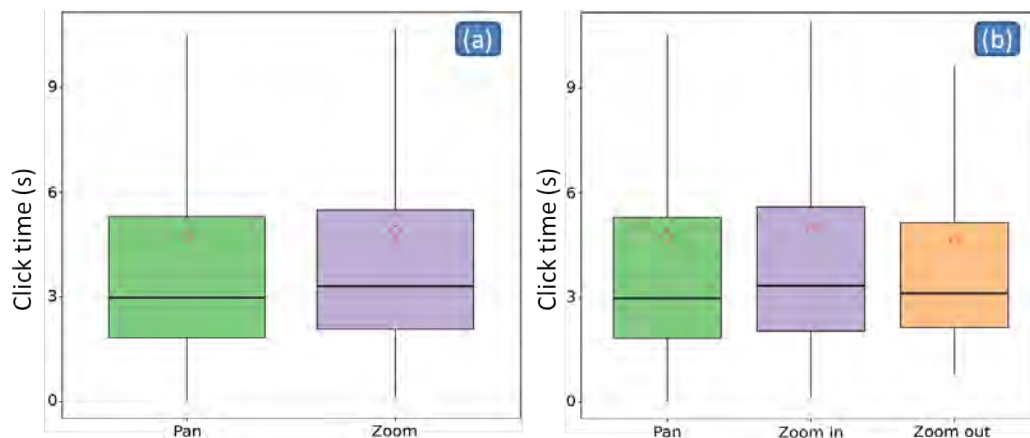
2.3 Procedure

The FogDetector survey follows a within-subject design, *i.e.* the participant carries out the task for all variable conditions, and even several times for each variable condition. The main variable of the survey is the interaction techniques performed before the task: either a zoom or a pan. Each of the techniques can be decomposed into several sub-techniques. A zoom can either be a zoom-in or a zoom-out and can cover a large scale gap (4 zoom levels difference) or a small scale gap (2 zoom levels difference). To balance the number of interactions, panning is also divided into two sub-techniques, panning at a large scale (zoom level 17) and panning at a small scale (zoom level 12).

In order to address (H_4), the other variable in the survey is the pan-scalar map used to perform the task. Three maps were selected: PLAN IGN, OPENSTREETMAP, and IGN CLASSIC. IGN CLASSIC is composed of scanned paper topographic maps produced by IGN. PLAN IGN is a new map designed by IGN as a pan-scalar map with a consistent style across scales to reduce disorientation. Finally, OPENSTREETMAP is the default OpenStreetMap pan-scalar map.

As disorientation can be caused by a loss of visual cues, we selected a task that requires the use of visual cues to be completed. We designed a recall task [1] where a target is shown on a map for 30 seconds. Then, the map switches to a different map view, and an animation navigates from this map view to the area of the target, with one of the sub-techniques (zoom or pan). Then, the user has 60 seconds to click on the location where they recall the target. The interaction is passive to make sure the participants directly go to the good view, at the cost of realism. The success of the task is assessed by the time of completion. The 31 targets

¹ <https://github.com/LostInZoom/lostinzoom-experiments>



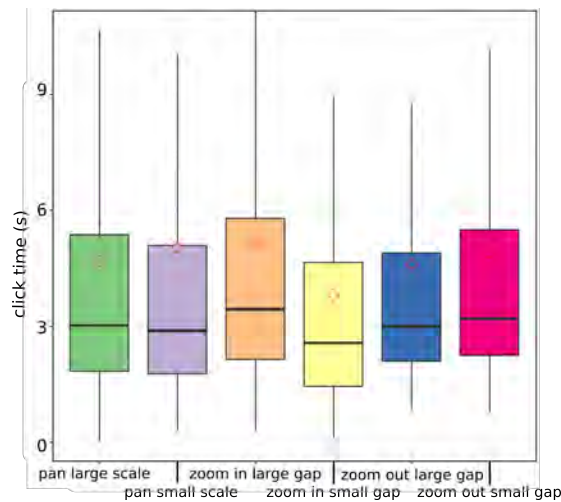
■ **Figure 2** (a) Box plot of the click time according to the interaction (pan or zoom). (b) Box plot of the click time according to the type of zoom. The red diamond shape shows the mean value.

are either buildings, crossroads, or points of interest (e.g. a specific symbol on the map, the centre of a lake, etc.) (Figure 1). The procedure is composed of a training phase, 2 blocks of 8 trials where all the sub-techniques are proposed to the participant with the same map, and 3 blocks of 6 trials where the three map designs are proposed, in order to compare them.

3 Results

Figure 2a shows the results of task completion time for pan vs zoom interactions. The difference of median time is 312 milliseconds (2.921 s for pan, and 3.233 s for zoom). A one-way repeated measures ANOVA was performed to compare the effect of general interaction techniques (pan or zoom) on task completion time. The ANOVA revealed that there was no statistically significant difference in mean completion time between the two groups ($F(1, 158) = [0.676], p = 0.41$). These results invalidate (H_1) in general, as the time difference between all types of zoom and all types of pan is not statistically significant. Figure 2b shows the results for the completion time difference between pan, zoom-in and zoom-out. The mean time for zoom-in is 4.866 s (*median* = 3.257), which makes a 218 milliseconds difference with pan (336 ms for the median). The mean time for zoom-out is 4.573 s (*median* = 3.104), which makes a -75 milliseconds difference with pan (183 ms for the median). A one-way repeated measures ANOVA was performed to compare the effect of interaction techniques (pan, zoom in, or zoom out) on task completion time. The ANOVA revealed that there was not a statistically significant difference in completion time between two groups, though the p value is not too important ($F(2, 346) = [1.747], p = 0.17$). (H_2) is not validated by this result but the difference we can observe is not the one we expected, because if zooming in seems to cause more disorientation than panning, zooming out appears to be an easier task.

Figure 3 shows the box plot of the completion times for each type of precise interaction: pan at small scale, pan at large scale, zoom in with a small scale gap, zoom in with a large scale gap, zoom out with a small scale gap and zoom out with a large scale gap. Results about (H_3): A one-way repeated measures ANOVA was performed to compare the effect of the six precise interaction techniques on task completion time. The ANOVA revealed that there was a statistically significant difference in completion time between at least two groups ($F(5, 790) = [5.550], p = 0.00005$). A post-hoc Tukey test found that the mean completion time value was significantly different between the two zoom-in interactions. These results



■ **Figure 3** Box plot of the click time according to the precise interaction used in the trial.

confirm (H_3) for zoom-in, with a completion time significantly higher when the scale gap is large. But the hypothesis is not confirmed for zoom-out where no significant difference was found in the Tukey test. These results also partially validate (H_1) and (H_2). Indeed, the mean time difference of 506 milliseconds between zoom-in large gap and pan large scale is statistically significant, but the difference between both zoom-out interactions and the others are never statistically significant.

To verify (H_4), we also looked at the differences in time completion between OpenStreetMap and both IGN maps. A one-way repeated measures ANOVA was performed to compare the effect of the map on task completion time. The ANOVA revealed that there was no statistically significant difference in mean completion time between the three groups ($F(2, 316) = [1.020], p = 0.36$).

4 Conclusion and future work



To conclude, the FogDetector survey allows a first measure of the disorientation occurring during a zoom interaction in a pan-scalar map, more than with a pan interaction. But the difference is only measured as significant when the scale change is large during a zoom-in. Surprisingly, the survey does not show any significant difference for zoom-out, probably due to the visual complexity of the maps displayed after our zoom-in interactions, compared to the maps displayed after a zoom-out. The survey shows no influence of the three maps used in the survey, so disorientation will not be significantly reduced just by adjusting the multi-scale style and content of the map.

To go further, the survey confirmed that the impact of disorientation was generally comparable to the duration of the pre-attentive phase of visual search [8], i.e. the time before we are able to focus our gaze on some target, and future studies should use quicker tasks, where pre-attention is even more crucial. One of the problems with our protocol is the fact that the interaction was passive, and we would like to perform a survey with active explorations from the user. Another direction is to couple a recall task with eye-tracking to measure cognitive load [15], as disorientation can be seen as a cause of cognitive load. Finally, as our final goal is to reduce disorientation, this survey is a first effort to work on pan-scalar design [5], but we know that just changing the style will not be sufficient.

References

- 1 Ann-Kathrin Bestgen, Dennis Edler, Kristina Müller, Patrick Schulze, Frank Dickmann, and Lars Kuchinke. Where Is It (in the Map)? Recall and Recognition of Spatial Information. *Cartographica*, 52(1):80–97, 2017. doi:10.3138/cart.52.1.3636.
- 2 Paul Dudchenko. *Why People Get Lost: The Psychology and Neuroscience of Spatial Cognition*. Oxford University Press, Oxford, 1st edition edition, September 2010.
- 3 Niklas Elmqvist, Andrew V. Moere, Hans-Christian Jetter, Daniel Cernea, Harald Reiterer, and T. J. Jankun-Kelly. Fluid interaction for information visualization. *Information Visualization*, 10(4):327–340, 2011. doi:10.1177/1473871611413180.
- 4 Pablo Fernández Velasco and Roberto Casati. Subjective disorientation as a metacognitive feeling. *Spatial Cognition & Computation*, 20(4):281–305, October 2020. doi:10.1080/13875868.2020.1768395.
- 5 Maïeul Gruget, Guillaume Touya, and Ian Muehlenhaus. Missing the city for buildings? a critical review of pan-scalar map generalization and design in contemporary zoomable maps. *International Journal of Cartography*, 2023. doi:10.1080/23729333.2022.2153467.
- 6 Mark Harrower and Benjamin Sheesley. Designing Better Map Interfaces: A Framework for Panning and Zooming. *Transactions in GIS*, 9(2):77–89, 2005. doi:10.1111/j.1467-9671.2005.00207.x.
- 7 Susanne Jul and George W. Furnas. Critical Zones in Desert Fog: Aids to Multiscale Navigation. In *Proceedings of the 11th Annual ACM Symposium on User Interface Software and Technology*, UIST '98, pages 97–106, New York, NY, USA, 1998. ACM. event-place: San Francisco, California, USA. doi:10.1145/288392.288578.
- 8 Bela Julesz. A theory of preattentive texture discrimination based on first-order statistics of textons. *Biological Cybernetics*, 41(2):131–138, August 1981. doi:10.1007/BF00335367.
- 9 Daniel R. Montello. Geographic orientation, disorientation, and misorientation: a commentary on Fernandez Velasco and Casati. *Spatial Cognition & Computation*, 20(4):306–313, 2020. doi:10.1080/13875868.2020.1767105.
- 10 Matthew Plumlee and Colin Ware. Zooming, multiple windows, and visual working memory. In *Proceedings of the Working Conference on Advanced Visual Interfaces*, AVI '02, pages 59–68, New York, NY, USA, May 2002. Association for Computing Machinery. doi:10.1145/1556262.1556270.
- 11 Irvin Rock, Christopher M Linnett, Paul Grant, and Arien Mack. Perception without attention: Results of a new method. *Cognitive Psychology*, 24(4):502–534, October 1992. doi:10.1016/0010-0285(92)90017-V.
- 12 Robert E. Roth. Interactive maps: What we know and what we need to know. *Journal of Spatial Information Science*, 6:59–115, 2013. URL: <http://josis.org/index.php/josis/article/view/105>.
- 13 Guillaume Touya, Maïeul Gruget, and Ian Muehlenhaus. Where Am I Now? Modelling Disorientation in Pan-Scalar Maps. *ISPRS International Journal of Geo-Information*, 12(2):62, February 2023. doi:10.3390/ijgi12020062.
- 14 Peter U. Tse. Mapping visual attention with change blindness: new directions for a new method. *Cognitive science*, 28(2):241–258, 2004. doi:10.1016/j.cogsci.2003.12.002.
- 15 Johannes Zagermann, Ulrike Pfeil, and Harald Reiterer. Measuring Cognitive Load using Eye Tracking Technology in Visual Computing. In *Proceedings of the Sixth Workshop on Beyond Time and Errors on Novel Evaluation Methods for Visualization*, BELIV '16, pages 78–85, New York, NY, USA, October 2016. Association for Computing Machinery. doi:10.1145/2993901.2993908.

An Interpretable Index of Social Vulnerability to Environmental Hazards

Joseph V. Tuccillo  

Oak Ridge National Laboratory, TN, USA

Abstract

Index-based measures of social vulnerability to environmental hazards are commonly modeled from composites of population-level risk factors. These models overlook individual context in communities' experiences of environmental hazards, producing metrics that may hinder spatial decision support for mitigating and responding to hazards. This paper introduces an interpretable, high-resolution model for generating an individual-oriented social vulnerability index (IOSVI) for the United States built on synthetic populations that couples individual and social determinants of vulnerability. The IOSVI combines an individual vulnerability index (IVI) that ranks individuals in an area's synthetic population based on intersecting risk factors, with a social vulnerability index (SVI) based on the population's cumulative distribution of IVI scores. Interpretability of the IOSVI procedure is demonstrated through examples of national, metropolitan, and neighborhood (census tract) level spatial variation in index scores and IVI themes, as well as an exploratory analysis examining risk factors affecting a specific sub-population (military veterans) in areas of high social and environmental vulnerability.

2012 ACM Subject Classification Information systems → Geographic information systems

Keywords and phrases Social Vulnerability, Environmental Hazard, Synthetic Population, Census, Veteran

Digital Object Identifier 10.4230/LIPIcs.GIScience.2023.74

Category Short Paper

Funding Notice: This work is sponsored by the US Department of Veterans Affairs. This manuscript has been authored by UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the US Department of Energy (DOE). The US government retains and the publisher, by accepting the article for publication, acknowledges that the US government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for US government purposes. DOE will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (<http://energy.gov/downloads/doe-public-access-plan>).

1 Introduction

Measuring and monitoring communities' social vulnerability to environmental hazards is a key consideration for planning decision support [19]. Social vulnerability (SV) broadly describes a population's collective potential for impacts from adverse events and circumstances [1], including natural hazards [3], technological hazards [10], and social determinants of health [18]. Measuring SV is complex, encompassing many conditions of everyday life, including demographics, socioeconomic status, living arrangement, housing, and mobility, that contribute to differential risk of harm or loss for communities exposed to a hazard [25].

Modeling SV often involves distilling multiple risk factors into composite indices that provide high-level characterizations of SV in an area of interest [5]. SV indices serve as entry points for more detailed analysis, including through descriptive characterizations of population risk [23] and field observations. SV modeling typically combines population-level variables (e.g., percentage in poverty, median age) into a composite score using dimensionality



© Joseph V. Tuccillo;
licensed under Creative Commons License CC-BY 4.0

12th International Conference on Geographic Information Science (GIScience 2023).

Editors: Roger Beecham, Jed A. Long, Dianna Smith, Qunshan Zhao, and Sarah Wise; Article No. 74; pp. 74:1–74:6

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

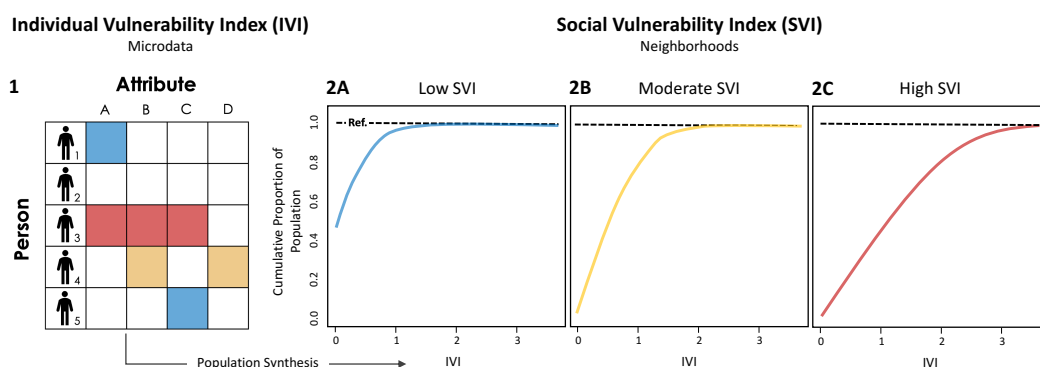


Figure 1 Conceptual illustration representing coupled Individual Vulnerability Index (IVI) scoring (Panel 1) and Social Vulnerability Index (SVI) scoring (Panels 2A - 2C).

reduction [2] or hierarchical aggregation methods [6]. A downside of these methods is that they exclusively compare areas' populations, thereby overlooking characteristics of residents who are likely to be directly impacted by hazards. In this way, ignoring individual risk factors in index construction poses challenges for the interpretability – and therefore actionability – of population-level SV models [11, 20].

This paper introduces an initial model for constructing an individual-oriented SV index (IOSVI) for the United States. The IOSVI supports greater interpretation of how individual vulnerability contributes to SV at the high spatial resolution of census tracts (1200 - 8000 people). The IOSVI is estimated from virtual or *synthetic* populations generated on public-use census microdata. Relative to model interpretability paradigms [13], the mechanisms for producing IOSVI are both *transparent* – built on open data and easily demonstrable [4] – and *decomposable* in that any one individual's level of vulnerability may be understood within the context of community SV, and vice-versa. As a result, the IOSVI is also interrogable, lending to *post-hoc analysis* (exploration, visualization) of individual/community characteristics within the context of SV.

2 The Individual-Oriented Social Vulnerability Index (IOSVI)

Individual “function-based” vulnerability is measured from intersecting *risk factors* that describe a person's daily sensitivities and may compound to affect their health and safety in adverse circumstances [17]. Common risk factors are tied to reduced socioeconomic status, living arrangement and age sensitivity, cultural sensitivity, and issues of housing and mobility [6]. Building upon the concept of function-based vulnerability, the IOSVI has two components: an Individual Vulnerability Index (IVI), which is a tabulation of the number of risk factors attributed to a member of an area's population, and the Social Vulnerability Index (SVI), which describes the cumulative distribution of the IVI within the area's population. The IVI can be a simple count of risk factors, but it can also be a hierarchical or weighted tabulation in instances where different categories of risk factors are of interest.

Figure 1 displays the cumulative distribution of resident IVI, ranked from low to high, for three hypothetical neighborhood areas. SVI can be measured as the difference in the areas under the curve (AUCs) between an area's observed cumulative IVI distribution and a reference distribution based on a hypothetical population in which no individuals are characterized by the risk factors of interest. A larger observed AUC corresponds to lower SVI, since 100% of the cumulative proportion of population occurs at a low IVI score threshold

(panel 2A). The reference AUC, which equates to the total number of possible risk factors, is included in the SVI computation to ensure proper directionality of the scores (i.e., low scores → low social vulnerability; high scores → high social vulnerability).

3 Methods

The IOSVI was developed through Oak Ridge National Laboratory's (ORNL) UrbanPop project. UrbanPop uses a regularized version of the Iterative Proportional Fitting (IPF) algorithm [16] to produce attribute-rich synthetic populations matched to large volumes of variables from the American Community Survey (ACS), the U.S. Census Bureau's primary intercensal product [22]. Toward developing IOSVI, this enables creating customizable representations of individuals in census tracts across the United States with respect how they embody risk factors contributing to SV. A series of 30 replicate synthetic populations for the development of IOSVI were produced for the United States, constrained on the ACS 2019 5-Year Estimates by adapting 14 risk factors identified by the Center for Disease Control and Prevention's (CDC) [6] at the individual level: **socioeconomic variables** including income below poverty, unemployment, and less than high school educational attainment; **living arrangement variables** including age over 65, age under 18, single-adult caregiver households, and disability status; **cultural sensitivity variables** including racial/ethnic minority status and limited English proficiency; and **housing/mobility variables** including multi-unit structures, mobile homes, household crowding, group quarters residency, and lack of a personal vehicle. SVI was computed at the tract level for each synthetic population replicate, following the method presented in Section 2. The final IOSVI was then computed as the Monte Carlo estimate (mean) of the replicate SVIs.

4 Illustrations

4.1 Visualizing Spatial Variation in the Social Vulnerability Index

Figure 2 maps the spatial distribution of IOSVI across the continental United States (panel A) and demonstrates visual interpretation of IOSVI for a portion of the Houston-The Woodlands, Sugar Land, TX Metropolitan Statistical Area (Houston MSA) (panels B, C). In panel C, each census tract's SVI score breaks down to four themes, identified via Multiple Correspondence Analysis (MCA), that describe the blend of risk factors best describing each profile of individual characteristics within the Houston MSA's synthetic population.

4.2 Examining Coupled Individual-Social and Environmental Vulnerability

A case study of intersecting individual, social, and physical (environmental) vulnerabilities was developed to demonstrate post-hoc exploratory analysis of the UrbanPop model underlying IOSVI. This example, developed for the Houston MSA, concerns a specific sub-population, U.S. military veterans. In emergency and disaster scenarios, veterans may experience pronounced problems of housing and livelihood recovery that impact mental and physical health [9, 15]. ACS provides indicators of veteran status for individuals age 25 and over, which were included as constraints for the UrbanPop model, then used to produce an indicator of veteran status in the synthetic population.

The exploratory analysis examines the association between veteran risk factors used to compute the IVI in combination with residency in high IOSVI - high physical vulnerability census tracts. Physical vulnerability was represented by a composite measure of annual

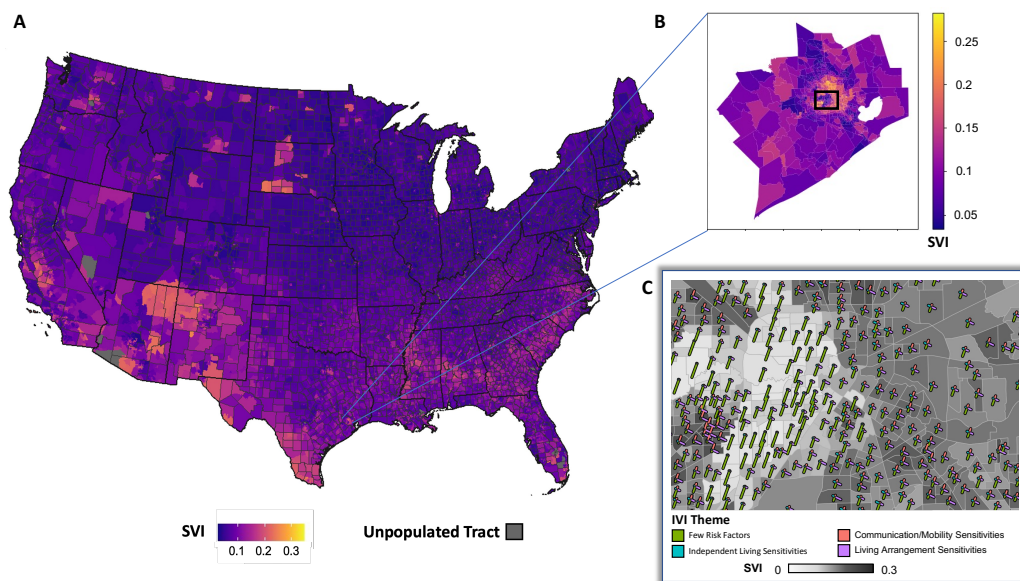


Figure 2 Mapping IOSVI to census tracts for the continental United States (Panel A), regionally (Panel B), and in neighborhood context relative to individual vulnerability index (IVI) themes (Panel C) (glyph plot methodology adapted from [14]).

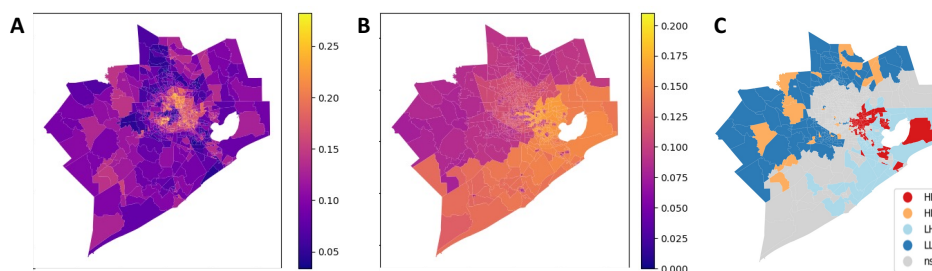


Figure 3 Overview of exploratory analysis of Individual-Oriented Social Vulnerability Index (IOSVI) for the Houston MSA. Panel A: IOSVI; Panel B: Composite annualized frequency score (AFS) of natural hazards from National Risk Index (NRI); Panel C: Bivariate Local Moran's I (BVMI) clusters for AFS by IOSVI. Abbreviations: H = High, L = Low, NS = Not Significant.

frequency of 18 hazard types for each census tract in the U.S. from the Federal Emergency Management Agency's (FEMA) National Risk Index (NRI) [26]. Bivariate Local Moran's I [12] was used to identify clusters of tracts in Houston with differing high (H) and low (L) combinations of social and physical vulnerability. For the subset of veterans in each tract's population, the association between IVI risk factors and residency in a high social vulnerability/high physical vulnerability (HH) tracts was evaluated against all other tract types, including non-significant (NS), using multinomial log-linear regression, a method specialized for dependent variables with multiple categorical labels [24].

Figure 4 reveals that risk factors frequently linked to uneven disaster recovery such as minority race/ethnicity and poverty [7, 8] were consistently associated (coefficient < 0) with veterans living in HH tracts in the Houston MSA, as were limited mobility (no car), less than high school education, disability status, single-caregiver households, and housing density (10+ units in structure).

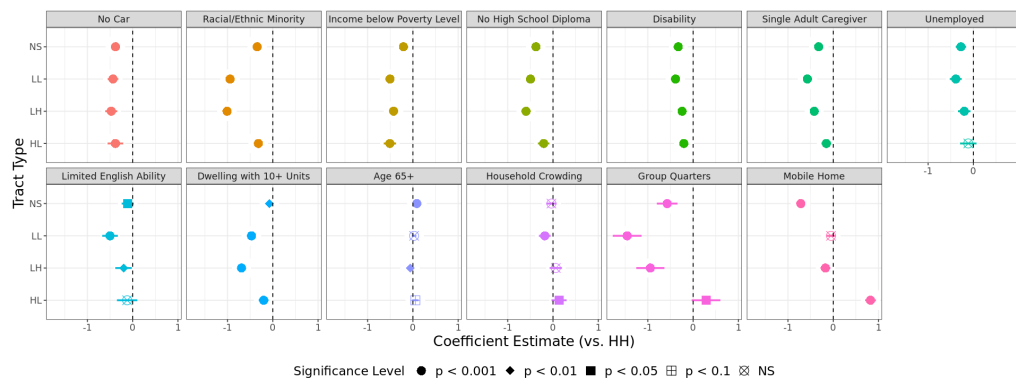


Figure 4 Multinomial regression coefficients and significances, BVMI (AFS by IOSVI) tract types by IOSVI risk factors for U.S. military veterans age ≥ 25 (base: high IOSVI - high AFS tracts). Abbreviations: H = High, L = Low, NS = Not Significant.

5 Conclusion and Outlook

Preliminary development of the individual-oriented social vulnerability index (IOSVI) on synthetic populations suggests enhanced interpretability (transparent, decomposable, interrogable) over existing methods that measure social vulnerability on population-level data alone.

Illustrations provided for the Houston MSA (Sections 4.1, 4.2) demonstrate the various ways that IOSVI may be examined: in national context, within census tracts, and for specific sub-populations (e.g., U.S. military veterans) in areas of high social/environmental vulnerability. Together, these insights may benefit more direct understanding of the spatial planning and policy interventions appropriate for addressing natural and technological hazards, as well as environmental determinants of health, at the neighborhood and community scales.

The primary limitation of IOSVI is the complexity of its design: large compute resources are required to generate and attribute synthetic populations, as well as estimate SV scores. This could be alleviated by developing an analytics platform for processing user requests and facilitating exploratory analysis, as well as supporting index development for custom geographies.

References

- 1 W Neil Adger. Vulnerability. *Global environmental change*, 16(3):268–281, 2006.
- 2 Susan L Cutter, Bryan J Boruff, and W Lynn Shirley. Social vulnerability to environmental hazards. *Social science quarterly*, 84(2):242–261, 2003.
- 3 Oronde Drakes and Eric Tate. Social vulnerability in a multi-hazard context: a systematic review. *Environmental research letters*, 2022.
- 4 David M Eddy, William Hollingworth, J Jaime Caro, Joel Tsevat, Kathryn M McDonald, and John B Wong. Model transparency and validation: a report of the ispor-smdm modeling good research practices task force–7. *Medical Decision Making*, 32(5):733–743, 2012.
- 5 Alexander Fekete. Spatial disaster vulnerability and risk assessments: challenges in their quality and acceptance. *Natural hazards*, 61:1161–1178, 2012.
- 6 Barry E Flanagan, Edward W Gregory, Elaine J Hallisey, Janet L Heitgerd, and Brian Lewis. A social vulnerability index for disaster management. *Journal of homeland security and emergency management*, 8(1), 2011.
- 7 Alice Fothergill, Enrique GM Maestas, and JoAnne DeRouen Darlington. Race, ethnicity, and disasters in the united states: A review of the literature. *Disasters*, 23(2):156–173, 1999.

- 8 Alice Fothergill and Lori A Peek. Poverty and disasters in the united states: A review of recent sociological findings. *Natural hazards*, 32(1):89–110, 2004.
- 9 June L Gin, Claudia Der-Martirosian, Christine Stanik, and Aram Dobalian. Roadblocks to housing after disaster: homeless veterans’ experiences after hurricane sandy. *Natural Hazards Review*, 20(3):04019005, 2019.
- 10 Ganlin Huang and Jonathan London. Mapping cumulative environmental effects, social vulnerability, and health in the san joaquin valley, california. *American journal of public health*, 102(5):830–832, 2012.
- 11 Brenda Jones and Jean Andrey. Vulnerability index construction: methodological choices and their influence on identifying vulnerable neighbourhoods. *International journal of emergency management*, 4(2):269–295, 2007.
- 12 Sang-Il Lee. Developing a bivariate spatial association measure: an integration of pearson’s r and moran’s i. *Journal of geographical systems*, 3:369–385, 2001.
- 13 Zachary C Lipton. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57, 2018.
- 14 Binbin Lu, Paul Harris, Martin Charlton, and Chris Brunsdon. The gwmodel r package: further topics for exploring spatial heterogeneity using geographically weighted models. *Geo-spatial Information Science*, 17(2):85–101, 2014.
- 15 Maria I Marshall, Linda S Niehm, Sandra B Sydnor, and Holly L Schrank. Predicting small business demise after a natural disaster: an analysis of pre-existing conditions. *Natural Hazards*, 79:331–354, 2015.
- 16 Nicholas N Nagle, Barbara P Battenfield, Stefan Leyk, and Seth Spielman. Dasymetric modeling and uncertainty. *Annals of the Association of American Geographers*, 104(1):80–95, 2014.
- 17 New York City Department of Health and Mental Hygiene. Vulnerable populations: A function-based vulnerability measure for the new york city region. https://www1.nyc.gov/assets/doh/downloads/pdf/em/regional_hazards_vulnerability_measures.pdf, 2013.
- 18 Alessandro Paro, J Madison Hyer, Adrian Diaz, Diamantis I Tsilimigras, and Timothy M Pawlik. Profiles in social vulnerability: the association of social determinants of health with postoperative surgical outcomes. *Surgery*, 170(6):1777–1784, 2021.
- 19 Samuel Rufat, Eric Tate, Christopher T Emrich, and Federico Antolini. How valid are social vulnerability models? *Annals of the American Association of Geographers*, 109(4):1131–1153, 2019.
- 20 Seth E Spielman, Joseph Tuccillo, David C Folch, Amy Schweikert, Rebecca Davies, Nathan Wood, and Eric Tate. Evaluating social vulnerability indicators: criteria and their application to the social vulnerability index. *Natural hazards*, 100:417–436, 2020.
- 21 Joseph Tuccillo and James Gaboardi. Likeness: a toolkit for connecting the social fabric of place to human dynamics. Technical report, Oak Ridge National Lab.(ORNL), Oak Ridge, TN (United States), 2022.
- 22 Joseph Tuccillo, Robert Stewart, Amy Rose, Nathan Trombley, Jessica Moehl, Nicholas Nagle, and Budhendra Bhaduri. Urbanpop: A spatial microsimulation framework for exploring demographic influences on human dynamics. *Applied Geography*, 151:102844, 2023.
- 23 Joseph V Tuccillo and Seth E Spielman. A method for measuring coupled individual and social vulnerability to environmental hazards. *Annals of the American Association of Geographers*, 112(6):1702–1725, 2022.
- 24 W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, New York, fourth edition, 2002. ISBN 0-387-95457-0. URL: <https://www.stats.ox.ac.uk/pub/MASS4/>.
- 25 Benjamin Wisner, Piers M Blaikie, Piers Blaikie, Terry Cannon, and Ian Davis. *At risk: Natural Hazards, People’s Vulnerability, and Disasters*. Psychology Press, 2004.
- 26 Casey Zuzak, Matthew Mowrer, Emily Goodenough, Jordan Burns, Nicholas Ranalli, and Jesse Rozelle. The national risk index: establishing a nationwide baseline for natural hazard risk in the us. *Natural Hazards*, 114(2):2331–2355, 2022.

Power of GIS Mapping: ATLAS Flood Maps 2022

Munazza Usmani¹ ✉ 🏠 

University of Trento, Italy

Fondazione Bruno Kessler, Trento, Italy

Hafiz Muhammad Tayyab Bhatti ✉

University of Punjab, Lahore, Pakistan

Francesca Bovolo ✉

Fondazione Bruno Kessler, Trento, Italy

Maurizio Napolitano ✉

Fondazione Bruno Kessler, Trento, Italy

Abstract

In this paper, we are introducing an efficient method based on the GIS technology, to design data immediate and analysis-ready mapping from open GIS and remote sensing data, vector and raster data into a single visualization to facilitate fast and flexible mapping, also referred to as ATLAS maps. The Google Earth Engine approach is used to pre-process the satellite data, while ArcGIS software is to integrate all the data layers. Since the ArcGIS software is included as a default dependency in GIS and remote sensing data, the proposed method provides a cross-platform and single-technology solution for handling flood mapping. For now, we conducted flood analysis using the latest open data for Pakistan and Nigeria countries, then elaborated on the advantages of each data for flood mapping with respect to inundated areas, rainfall analysis, and affected populations, health, and education facilities. Given a wide range of tasks that can benefit from the method, future work will extend the methodology to heterogeneous geodata (vector and raster) to support seamless and make it automatic interfaces.

2012 ACM Subject Classification Information systems → Geographic information systems

Keywords and phrases GIS, Disaster Mapping, Open Data, Geospatial Technology, Remote Sensing

Digital Object Identifier 10.4230/LIPICs.GIScience.2023.75

Category Short Paper

1 Introduction

Urban flooding is a serious issue in many cities across the world. They cause significant losses in terms of lives, possessions, buildings, and means of subsistence [2]. In 2022, Pakistan and Nigeria experienced severe flooding in various regions due to their topography and monsoon season, resulting in loss of life, property, and infrastructure. The use of Geographic Information Systems (GIS) and remote sensing techniques can be a valuable tool in mapping and analyzing the extent and impact of the flood. Also, social media and crowdsourcing applications have enabled real-time data collection from citizens in flood-prone areas. This data on integration can create situational awareness maps and inform emergency response efforts. Then this information can be used to model flood scenarios and assess the potential impact on communities and infrastructures. But when comes to integrating or fusing all data layers, it makes managing and monitoring floods very challenging. Therefore, in every flood-prone area, a thorough assessment of floods is crucial. In order to analyze flood inundation, this work shows a logical framework based on Sentinel-1 Synthetic Aperture

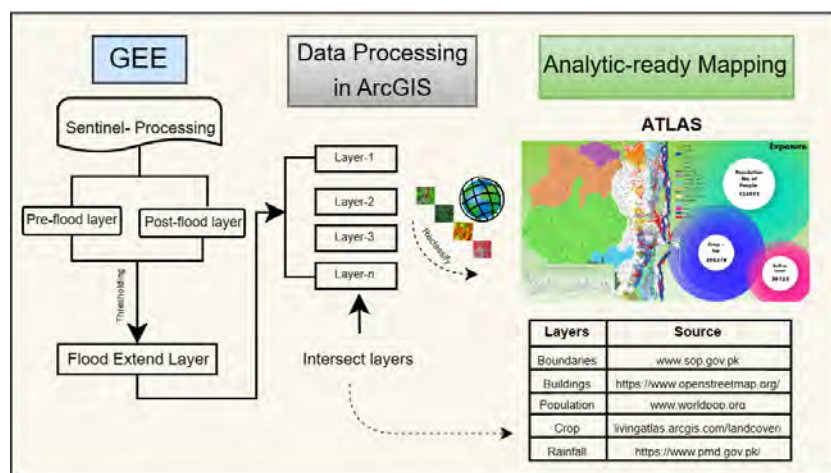
¹ corresponding author



Radar (SAR) data, which is then integrated with GIS data to produce flood ATLAS maps. A collection of different maps of the Earth or a particular area of the Earth is called an ATLAS. The maps in ATLAS display geographical characteristics, local topography, and political boundaries. Additionally, they provide information on a region's social, religious, and economic statistics [1]. The availability of open data and open-source GIS software has made flood ATLAS mapping more accessible to researchers, policymakers, and citizens. This has enabled the development of community-based flood management strategies and increased public participation in disaster risk reduction. By providing a comprehensive visualization of the flood impact, the ATLAS maps can also help to raise awareness among the public about the severity of the flood and the need for action to mitigate its effects [4]. This study first creates ATLAS flood maps of various factors in each province of Pakistan, including population, inundated areas, and rainfall analysis. In addition, it includes an ATLAS map of the torrents of Dera Ghazi (DG) Khan, depicting the water flow patterns in the region. DG Khan is a district located in the southern part of Punjab province, Pakistan, and is known for its rugged terrain and arid climate. The district is prone to flash floods, which can cause significant damage to the local communities, agriculture, and infrastructures in Punjab province. It includes a post-flood analysis showing inundated areas and affected populations, schools, and hospitals for each state of Nigeria. In the following, there is more detail of data and methodology in Section 2. The ATLAS maps for the flood-2022 are shown in Section 3 following the conclusion in Section 4.

2 Framework

Following Figure 1, a methodology for creating the ATLAS maps of flood disasters using remote sensing and openly available GIS data is shown. The main step in the process is to collect the necessary primary and secondary data. Primary data consists of Sentinel images downloaded from GEE, which provides free access to a wide range of remote sensing data. The data are obtained for August and October 2022, which is typically the peak monsoon season in Pakistan [3] and Nigeria, respectively.



■ **Figure 1** Framework for ATLAS Mapping.

The secondary data and modeling outputs from different data sources (Figure 1) are used to compile geospatial maps. Data used in the maps are also showing the importance of open data for flood mapping in emergencies.

The methodology's most crucial step is identifying the flooded areas by processing Sentinel-1 SAR data. The JavaScript programming language, which is directly integrated into the GEE interface, is deployed in this part. In addition to image processing, analysis of images, result visualizing, and result exporting, it covers command declaration tasks for importing image data to the platform. Pixel values from pre- and post-flood imagery have been obtained and compared for this investigation. A difference makes it easier to distinguish between pixels that represent areas that are flooded during the flood season and those that represent permanent water bodies like rivers and wetlands. The following describes the GEE flood mapping and inundated area computation methodology: the Sentinel-1 data package and study area boundary have been imported to GEE in the first step. In the next step, the time frame and sensor parameters were specified for this study. For Pakistan, the base period selected for flooding area comparison is from the 6th to the 27th of August, while for Nigeria is from the 13th of October to the 24th of October. By setting periods, the selection covers the specific season of the selected area. The "descending" pass direction and polarization "VV" of the sensor are the specified parameters. When the pixels that represented the flooded area were correctly identified, the extent of the flood was calculated. The key comparison is the variation between the photographs taken before and after the flood. The flood extent mask is made once the predetermined threshold has been applied, and the flood result has been improved. ArcGIS has been used to display the flood extent area data that was obtained from GEE. It is a desktop GIS tool that is cross-platform for browsing, editing, and analyzing geographic data.

Once the flood inundation layer is created, it is used to calculate the impact of the flood on populations, schools, and hospitals. This is done by overlaying the flood inundation layer with other openly available data sources. The impact analysis is carried out using ArcGIS software, which enabled the calculation of the affected population, schools, hospitals, and rainfall analysis.

The final step in the process is to create ATLAS maps that visualize the flood inundation and its impact. This is done using ArcGIS software and involved selecting an appropriate symbology and colors for the map and adding labels and legends. The ATLAS maps are designed to be visually appealing and easy to interpret with clear information. In conclusion, the methodology for creating the ATLAS maps of flood disasters in Pakistan and Nigeria involved collecting remote sensing and open source data, processing and analyzing the data using ArcGIS software, and finally visualizing the extent of the flood inundation and its impacts. This methodology can create similar ATLAS maps for other regions or different types of natural disasters.

3 ATLAS Maps

The following ATLAS maps give an overview of all the situations showing population, inundated area, and rainfall analysis. Starting from Pakistan, the situation in four provinces has been shown (from Figure 2 to Figure 6) with DG Khan torrent response in Figure 3, showing the situation of water levels/pressures on different water streams, rivers, and nullahs. Figure 7 glance into the overall status of Nigeria in terms of inundated areas, affected population, schools, and hospitals.

4 Conclusion

This flood ATLAS summarizes the findings of the post-flood assessment that took place in Pakistan and Nigeria in 2022. The 2022 floods in Pakistan, affected about 5 million people in total and an area of about 55,058 km^2 has been inundated. We also made an ATLAS map

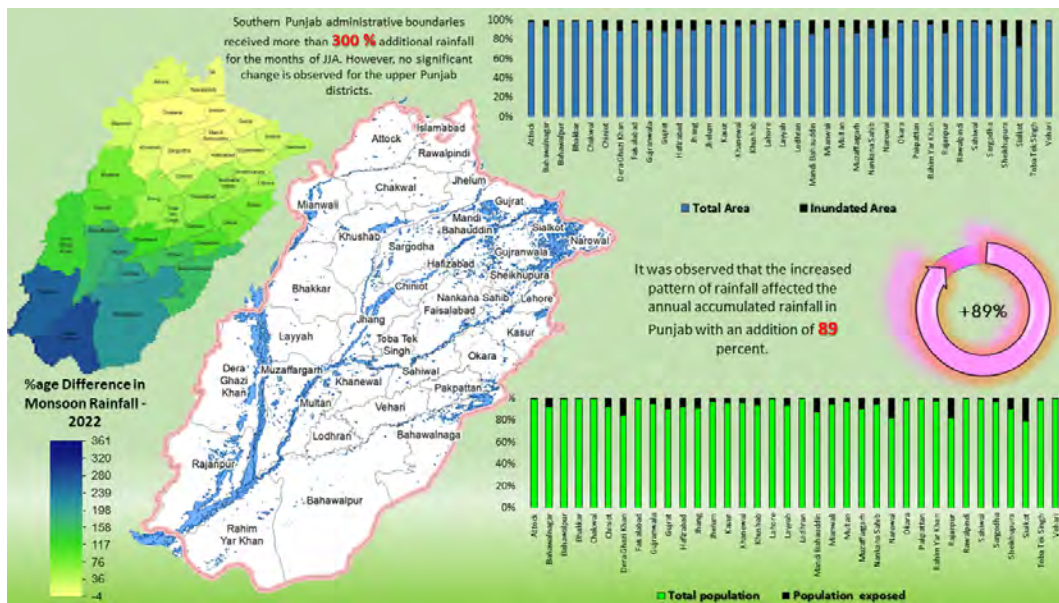


Figure 2 ATLAS Map for Punjab Province.

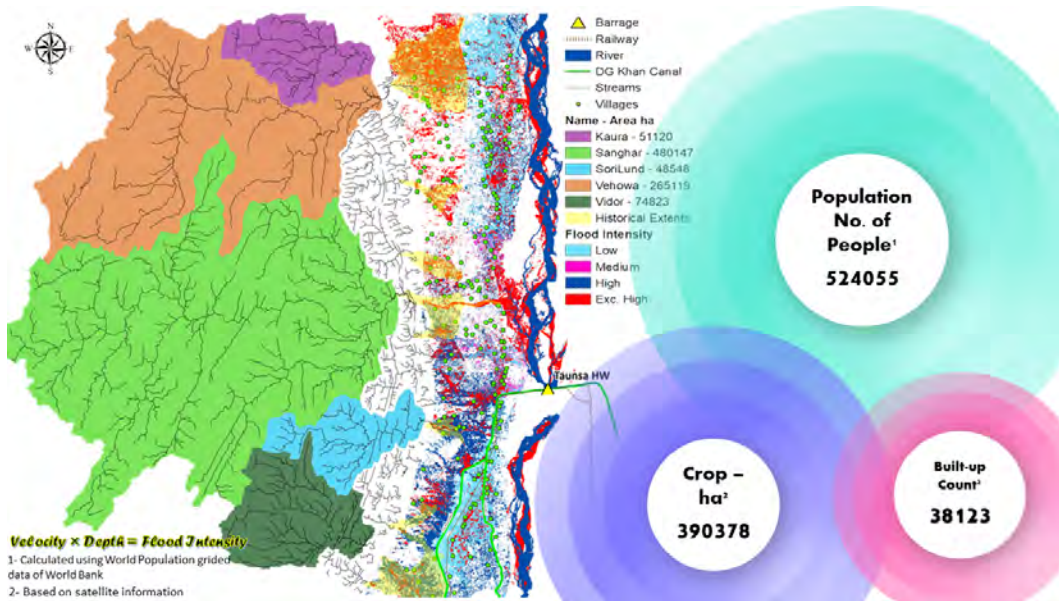


Figure 3 ATLAS Map DG Khan Torrent's Response.

of Nigeria's 2022 floods showing the affected population, flood extent, schools, and health facilities in each state. Future damage assessments of urban areas as well as the delineation and designation of existing floodplain boundaries, can all be updated by these maps.

References

- 1 Amanda. Briney. What is an atlas?, 2021.
- 2 Umar Lawal Dano, Abdul-Lateef Balogun, Abdul-Nasir Matori, Khmaruzzaman Wan Yusouf, Ismaila Rimi Abubakar, Mohamed Ahmed Said Mohamed, Yusuf Adedoyin Aina, and Biswajeet

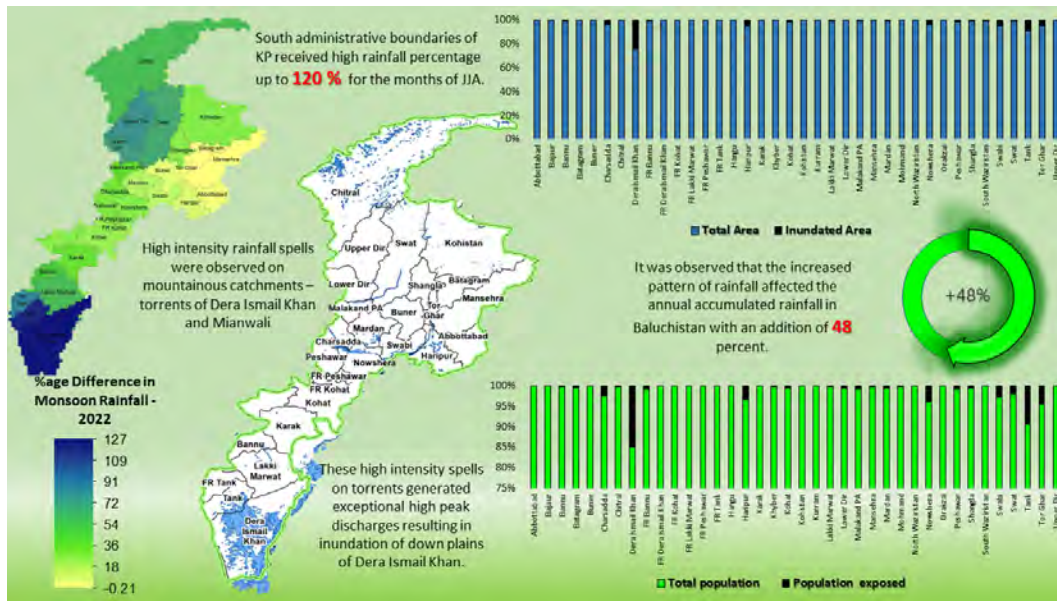


Figure 4 ATLAS Map for KPK Province.

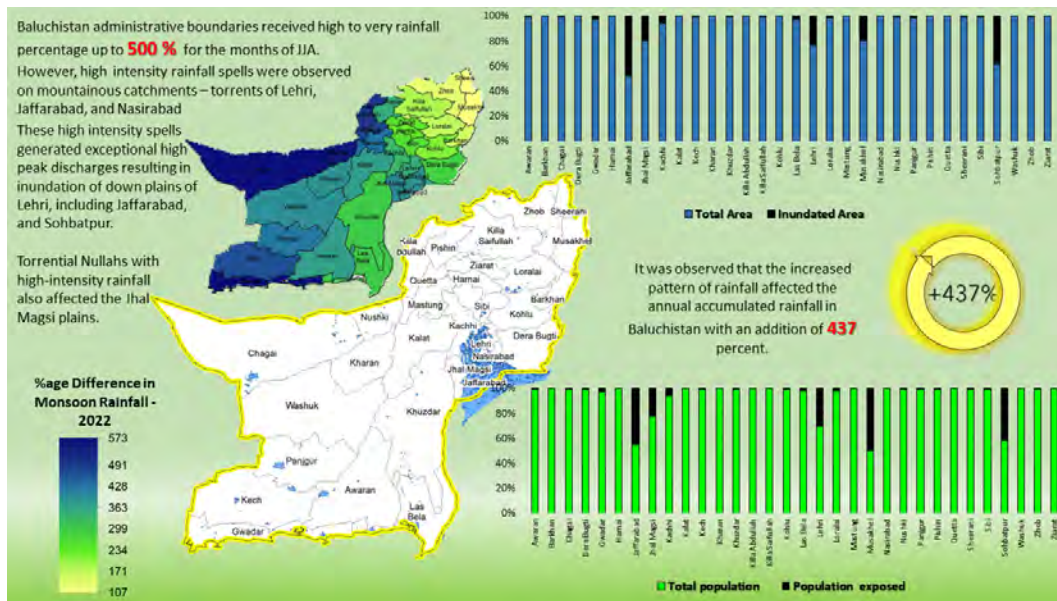


Figure 5 ATLAS Map for Baluchistan Province.

- Pradhan. Flood susceptibility mapping using gis-based analytic network process: A case study of perlis, malaysia. *Water*, 11(3):615, 2019.
- 3 Ejaz Hussaina, Serkan Urala, Abrar Malikb, and Jie Shana. Mapping pakistan 2010 floods using remote sensing data. In *Proceedings of the ASPRS Annual Conference, Milwaukee, WI, USA*, volume 15, 2011.
 - 4 Sidhant Ochani, Syeda Ilsa Aaqil, Abubakar Nazir, Fatima Binte Athar, Khushi Ochani, and Kaleem Ullah. Various health-related challenges amidst recent floods in pakistan; strategies for future prevention and control. *Annals of Medicine and Surgery*, 82:104667, 2022.

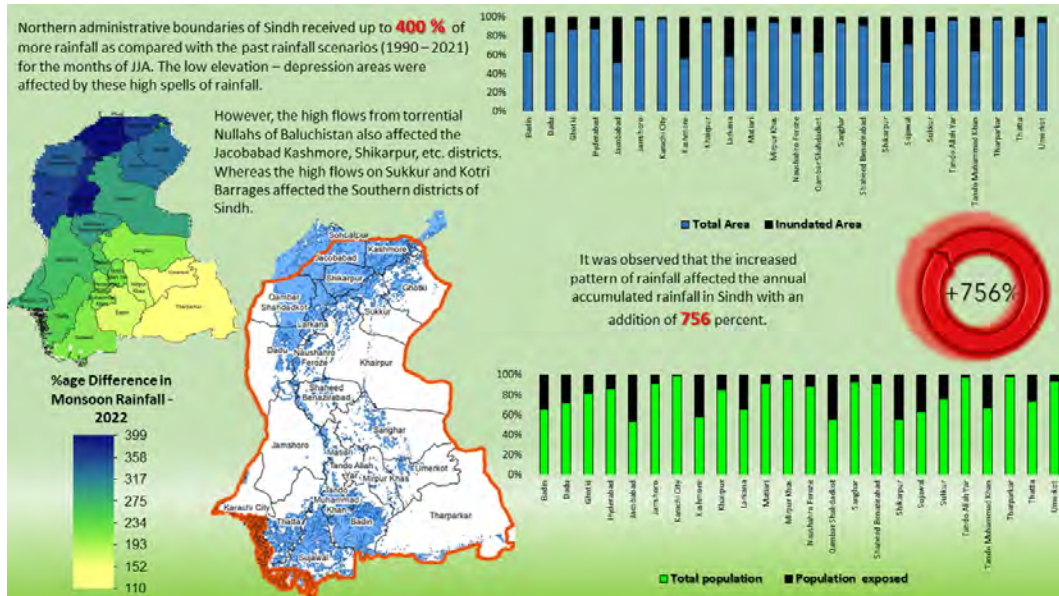


Figure 6 ATLAS Map for Sindh Province.

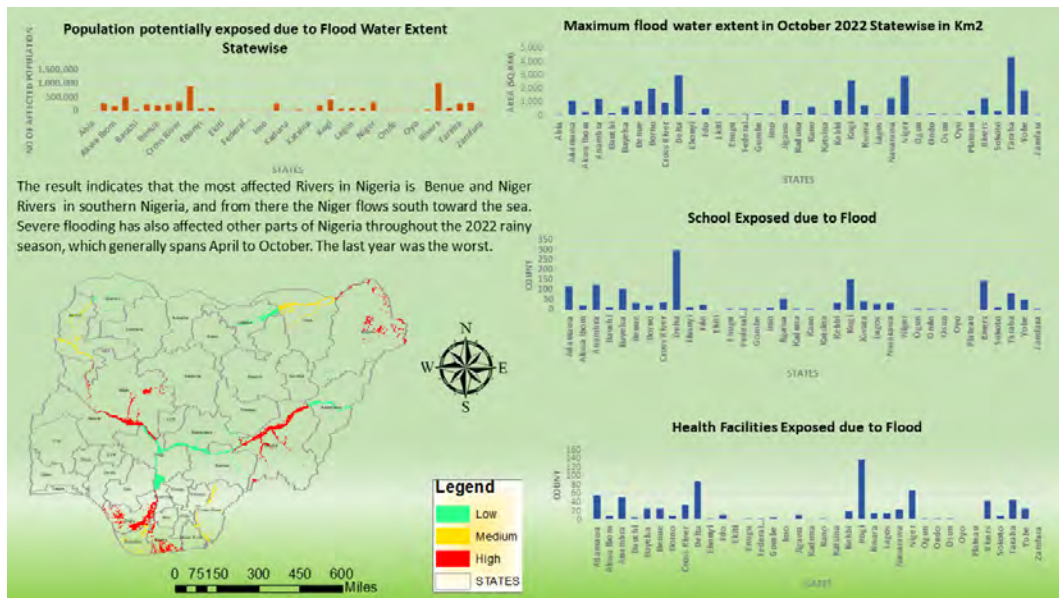


Figure 7 ATLAS Map for Nigeria.

A Data-Driven Decision-Making Framework for Spatial Agent-Based Models of Infectious Disease Spread

Emma Von Hoene¹ ✉ 

Department of Geography and Geoinformation Science, George Mason University, Fairfax, VA, USA

Amira Roess ✉ 

Department of Global and Community Health, George Mason University, Fairfax, VA, USA

Taylor Anderson ✉ 

Department of Geography and Geoinformation Science, George Mason University, Fairfax, VA, USA

Abstract

Agent-based models (ABMs) are powerful tools used for better understanding, predicting, and responding to diseases. ABMs are well-suited to represent human health behaviors, a key driver of disease spread. However, many existing ABMs of infectious respiratory disease spread oversimplify or ignore behavioral aspects due to limited data and the variety of behavioral theories available. Therefore, this study aims to develop and implement a data-driven framework for agent decision-making related to health behaviors in geospatial ABMs of infectious disease spread. The agent decision-making framework uses a logistic regression model expressed in the form of odds ratios to calculate the probability of adopting a behavior. The framework is integrated into a geospatial ABM that simulates the spread of COVID-19 and mask usage among the student population at George Mason University in Fall 2021. The framework leverages odds ratios, which can be derived from surveys or open data, and can be modified to incorporate variables identified by behavioral theories. This advancement will offer the public and decision-makers greater insight into disease transmission, accurate predictions on disease outcomes, and preparation for future infectious disease outbreaks.

2012 ACM Subject Classification Computing methodologies → Modeling methodologies

Keywords and phrases Agent-based model, geographic information science, disease simulation, COVID-19, agent behavior, mask use

Digital Object Identifier 10.4230/LIPICs.GIScience.2023.76

Category Short Paper

Funding National Science Foundation (Award #2030685 and #2109647).

Acknowledgements The survey used in this project was considered exempt by the George Mason University Institutional Review Board (IRB 1684418-3).

1 Introduction

In the twenty-first century, society has faced several infectious disease outbreaks, such as monkeypox, influenza, and the novel COVID-19 [2]. To mitigate these threats, decision-makers rely on epidemiological models for predicting outbreaks and assessing the impact of various interventions [1]. Compartmental models, while computationally efficient, struggle to capture heterogeneous populations, spatial interactions, and individual health behaviors - key drivers of disease trajectories [8]. An alternative approach is agent-based modeling (ABM), which explicitly represents behavior and interactions between individual “agents”

¹ corresponding author



and their environment. ABMs provide the flexibility to assign heterogeneous attributes and decision-making processes to each agent. In addition, spatial data can be used to represent the built environment, facilitating the representation of individual movements and interactions in space [8]. The ability to capture complex behaviors of individuals makes ABM an ideal approach for better understanding the spread of diseases and supporting decision-making.

Health behaviors, actions that affect disease transmission, are often overlooked or simplified in existing ABMs. This is mainly due to lack of data to inform agents' behavioral parameters, competing theories of health behavior to draw from, and institutional challenges limiting interdisciplinary collaboration among modelers and domain experts [5]. For instance, Perez and Dragicevic [14] model the spatial spread of disease without considering behavioral responses like staying home when sick. Other models impose behaviors on a set of agents without considering their individual characteristics, beliefs, or perceptions. For example, Li et al. [10] compare COVID-19 outcomes using scenarios with different percentages of randomly selected agents who are considered vaccinated. More complex representations of health behavior range from the use of social contagion of adopting behaviors along a social network [9], game theory or the rational choice model to inform adoption decisions [17], and fuzzy cognitive maps [12]. However, in general, most models of disease spread either ignore or incorporate health behaviors in an ad hoc manner without leveraging behavioral data or theories. Therefore, a more comprehensive framework that has the potential to leverage both data and theories for simulating health behaviors in ABMs of disease spread is needed.

Realistically incorporating human behaviors into ABMs of infectious disease spread demands a combination of data-driven and theoretical approaches. The objective of this study is to develop and implement a data-driven agent decision framework that improves the representation of health behaviors in geospatial ABMs of infectious disease spread. The framework is intended to leverage behavioral data and theories by implementing a logistic regression model with a geospatial ABM of infectious disease. This is demonstrated by simulating the spread of COVID-19 and mask-usage behavior among the undergraduate student population at George Mason University's (GMU) Fairfax campus in Fall 2021.

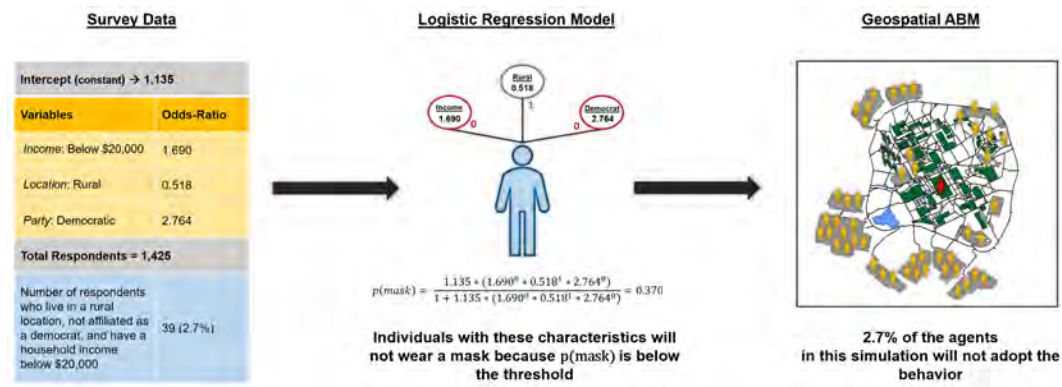
2 Methods

2.1 Data

Built Environment Data. The model environment consists of GMU's Fairfax campus, which was created using shapefiles obtained from GMU's open geoportal. Datasets capturing the buildings, parking lots, walkways, and water features were acquired and updated to reflect the built environment on campus in 2021. Within each building are sublocations in which agents interact, for example representing individual classrooms.

Data Informing University Patterns of Life. The agent schedules were generated using GMU's course data. All in-person courses offered during the Fall 2021 semester were collected from PatriotWeb, GMU's openly available database with a schedule of classes. 2,004 courses were collected, along with details on their course number, course section, meeting times, total seats, building and classroom. The data was processed into variables that directed the agents to enroll in courses that corresponded with their undergraduate program, attend class at the scheduled times, and travel to their course's designated building and classroom.

Health Behavior Data. This study utilizes data from a survey conducted in August 2021 that is representative of the United States [16]. The survey gathered information on factors influencing mask-wearing decisions when masks were optional in the U.S. 3,528 respondents



■ **Figure 1** A representation of how the agent decision-making framework works.

participated, providing socio-demographic details such as age, gender, ethnicity, income, political party affiliation, and rural or urban residency. The study exclusively focused on respondents under the age of 40, resulting in a total of 1,425 participants. Respondents were asked how often they decided to wear a mask. Odds ratios were calculated to capture the association between the socio-demographic variables and individuals who would wear a mask 3 or more times a week. The survey results indicated that only income, residence location, and political party affiliation were statistically significant variables for predicting mask use.

Disease Data for Calibration. The model is calibrated using data obtained from GMU’s Campus COVID-19 Data Archive. The dashboard includes COVID-19 case, testing, and vaccine data among the Mason community during the Fall 2021 semester, spanning from August 16, 2021 to December 17, 2021.

2.2 Agent Behavioral Framework

The proposed agent decision framework determines how agents in a geospatial ABM make decisions about whether or not to adopt a specific health behavior. A visual representation of how the agent decision-making framework is applied to the geospatial ABM in this study is presented in Figure 1. The framework is adapted from the methods presented by Durham and Casman [4] that uses a standard logistic regression model to calculate the probability that an agent will adopt a behavior based on four antecedents, defined by the Health Belief Model. The logistic regression model is expressed in terms of odd ratios, which are informed by survey data. We modify the methods presented by Durham and Casman [4] to include the variables that we have recognized from our survey data as important predictors of mask usage. These variables are outlined in Section 2.1 and are expressed in Equation 1:

$$p(\text{mask}) = \frac{OR_0 \cdot \prod_i OR_i^{X_i}}{1 + OR_0 \cdot \prod_i OR_i^{X_i}}, \quad i = 1, \dots, 3 \tag{1}$$

The values of $i = 1, \dots, 3$ denotes each of the independent variables that will be used, including income, residence location, and political party. OR_i is the value of the odds ratio. X_i is a binary variable that indicates the state of the independent variable where 1 is true and 0 is false. OR_0 is a constant probability when all X_i variables are low. The probability of behavior, $p(\text{mask})$, is a value from 0 to 1. A threshold is used to determine whether the individual adopts the behavior. When $p(\text{mask}) > \text{threshold}$, the individual will uptake the behavior, and if $p(\text{mask}) < \text{threshold}$, the individual will not uptake the given behavior.

2.3 Geospatial ABM

The geospatial ABM was developed and integrated with an agent decision framework that can collectively leverage both behavioral theories and data to realistically implement human health behavior during an infectious disease outbreak. The model aims to demonstrate this framework by simulating the spread of COVID-19 and mask-wearing behavior among the undergraduate population at GMU's Fairfax campus in Fall 2021.

Agent Characteristics and Scheduling. The model consists of agents representing undergraduate students who were registered for in-person classes in the Fall 2021 semester and enrolled full-time. University data informed whether the student agent lives on or off campus, as well as the elective and core courses associated with their undergraduate degree program. Based on the joint distributions found in the survey data, agents' demographic profiles are generated to include three binary characteristics: residence location (rural or otherwise), political affiliation (democratic or otherwise), and household income (below \$20,000 or otherwise). Given the odds ratios from the survey data and the demographic profile of the agent, the probability of mask-wearing from 0 to 1 for each agent is assigned. Additional agent characteristics that affect the disease transmission include health status, symptomatic status, and quarantine status. The length of the incubation and infectious periods for infectious agents is determined from a normal distribution informed by COVID-19 literature. All agents are assumed vaccinated as it was mandatory to be on campus during the Fall 2021 semester.

The model processes discrete time steps representing one second, captures weekly patterns (Monday-Friday from 7:15am to 11pm), and stops at the end of the semester. Agents follow their class schedule generated at the initialization of the model, beginning each day at their home, represented in the model as either a parking lot or a residential building on campus. They leave for their first class 15 minutes before it starts and travel to each class throughout the day, updating the building and sublocation where they are located. If an off-campus agent has no class following their previous one, they go to a student center or gym until the next class, while on-campus agents have a 50 percent chance of going to a student center/gym or returning to their dorm. After attending all their classes for the day, agents return home.

Disease Transmission. Since COVID-19 was simulated in the Fall 2021 semester, literature was used to define the parameters of the Delta variant. There is a probability that susceptible agents are exposed to the virus if they come in contact with an infected agent, known as the transmission rate, which is calibrated to 0.05 (see Initial Calibration). However, if the agent is wearing a mask, then the transmission rate is reduced by 50% [6], resulting in a transmission rate of 0.025. Additionally, all susceptible agents have a 0.02 off-campus transmission rate, a value obtained from other disease spread models on a university campus [6].

Once an agent becomes exposed, they remain in the exposed stage for an average of 4.41 days, after which they have an 85% chance of becoming symptomatic [18, 13]. Both symptomatic and asymptomatic agents stay in the infectious period for an average of 8 days [7]. However, this infectious period includes a pre-symptomatic stage of two days because individuals can spread the virus 48 hours before symptoms appear [15]. After the pre-symptomatic stage, symptomatic agents begin to quarantine, meaning they do not go to campus until their infectious period has ended, where asymptomatic agents continue to follow their schedules [15]. Individuals have immunity for 90 days, which implies that agents remain in the recovered period for that length of time since they cannot be re-infected [3].

Health Behavior Framework. At the initialization of the model, agents decide to wear a mask for the period of a semester based on their demographic profile. A logistic regression model for each agent determines based on the combination of their characteristics what the probability of mask use is from 0 to 1. Agents will choose to wear a mask if that value is greater than a 0.50 threshold, which may be calibrated in the future. The model currently does not incorporate individual learning, sensing, or prediction. Future work will include agent perception leading to dynamic health behaviors by incorporating different driving variables such as the perceived severity or perceived susceptibility of the disease.

Initialization. The model is implemented with Repast Simphony, a freely available Java-based modeling toolkit, and built upon a Repast Simphony program called RepastCity [11]. The model is currently not accessible online, but it will be made available in the future once the work is completed. Before the model begins running, one agent is selected to be infectious, and 3.5% of agents are set to be recovered. The model is initialized with 5,000 agents, which captures 37.5% of the estimated campus population during the Fall 2021 semester. While the model will be upscaled to include 13,500 agents in future work, for the current study, we limited the simulation to 5,000 agents to test the proposed agent decision-making framework.

Initial Calibration. GMU's Fall 2021 COVID-19 data archive reported 399 symptomatic cases throughout the semester with 100% mask use. Since we simulate roughly 37% of the campus population, we expect around 148 symptomatic cases in our model. We calibrate the model by modifying the transmission rate in the 100% mask-use scenario so that the number of infectious and symptomatic agents throughout the semester corresponds with the data.

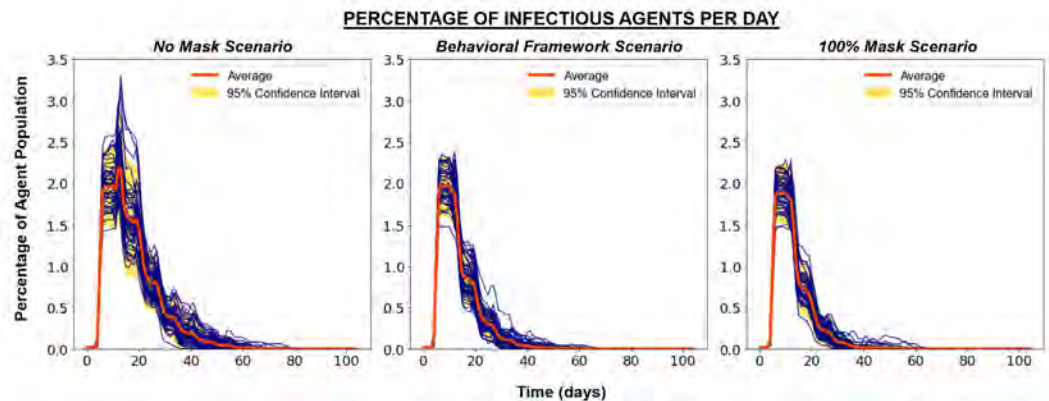
3 Results

To address the variation in model results due to randomness within ABM processes, the model was run 50 times for a duration of 105 days to represent a full 15-week semester at GMU. Future work will validate disease outcomes using data that was not used in model calibration prior to running scenarios and generating final results. However, we present initial results for three scenarios here: 1) no mask, 2) 100% mask usage, and 3) mask usage determined by the agent's behavioral framework, where the demographic profiles in the population determine the adoption of masks. The results are presented in Figure 2.

The preliminary results indicate that the number of cumulative cases was on average 253 for no mask-usage, 165 for mask-usage determined by the agent behavioral framework, and 148 for 100% mask-usage. No-mask usage results in a higher peak infection level and greater variation, whereas 100% mask usage leads to less variation across simulation runs and a lower percentage of infections. The findings from the behavioral framework scenario are comparable to those of 100% mask usage as most of the agent population chose to wear masks based on their demographic profile.

4 Discussion and Conclusion

Existing ABMs of infectious respiratory disease spread often overlook or oversimplify the complexities of human health behaviors. To address this gap, this study proposes a novel agent decision making framework with the potential to integrate data-driven and theoretical approaches. Limitations of this work include the use of national survey data to represent a university population. Future work may explore the use of open data that better corresponds



■ **Figure 2** The results for each tested scenario, showing the daily percentage of infectious agents.

with the study area. Although, the agents' health behaviors emerge as a function of each agent's demographic profile, they are unchanging. Future work will explore the effect of agent perceptions on behavior, enabling agents to adapt and respond to new situations. This study aims to advance ABMs of infectious disease spread by improving the representation of how humans respond to diseases, which will ultimately offer the public and decision-makers support with accurate predictions and intervention strategies for future outbreaks.

References



- 1 L. Berger and et al. Rational policymaking during a pandemic. *PNAS*, 118(4), 2021.
- 2 D.E. Bloom and D. Cadarette. Infectious disease threats in the twenty-first century: Strengthening the global response. *Frontiers in Immunology*, 10, 2019.
- 3 CDC. Reinfection: Clinical considerations for care of children and adults with confirmed covid-19, 2020. URL: <https://www.cdc.gov/coronavirus/2019-ncov/hcp/clinical-care/clinical-considerations-reinfection.html>.
- 4 D. Durham and E. Casman. Incorporating individual health-protective decisions into disease transmission models: a mathematical framework. *J R Soc Interface*, 9(68):562–570, 2012.
- 5 S. Funk, M. Salathé, and V. Jansen. Modelling the influence of human behaviour on the spread of infectious diseases: a review. *J R Soc Interface*, 7(50):1247–1256, 2010.
- 6 P.T. Gressman and J.R. Peck. Simulating covid-19 in a university environment. *Mathematical Biosciences*, 328, 2020.
- 7 X. He and et al. Temporal dynamics in viral shedding and transmissibility of covid-19. *Nature Medicine*, 26:672–675, 2020.
- 8 E. Hunter, B. Mac Namee, and J. Kelleher. A taxonomy for agent-based models in human infectious disease epidemiology. *JASS*, 20(3), 2017.
- 9 D.A. Levy and P.R. Nail. Contagion: a theoretical and empirical review and reconceptualization. *Genetic, social, and general psychology monographs*, 119(12):233–284, 1993.
- 10 Z. Li, J.L. Swann, and P. Keskinocak. Value of inventory information in allocating a limited supply of influenza vaccine during a pandemic. *PLOS ONE*, 13(10), 2018.
- 11 N. Malleson. Repastcity, 2012. URL: <https://code.google.com/archive/p/repastcity/>.
- 12 S. Mei and et al. Individual Decision Making Can Drive Epidemics: A Fuzzy Cognitive Map Study. *IEEE Transactions on Fuzzy Systems*, 22(2):264–273, 2014.
- 13 T. Ogata and et al. A low proportion of asymptomatic covid-19 patients with the delta variant infection by viral transmission through household contact at the time of confirmation in ibaraki, japan. *ISPRS International Journal of Geo-Information*, 4(3):192–196, 2022.

- 14 L. Perez and S. Dragicevic. An agent-based approach for modeling dynamics of contagious disease spread. *International journal of health geographics*, 8(1):1–17, 2009.
- 15 M. Reveil and Y.H. Chen. Predicting and preventing covid-19 outbreaks in indoor environments: an agent-based modeling study. *Scientific Reports*, 12, 2022.
- 16 A. Roess and et al. Predictors of firearm purchasing during the coronavirus pandemic in the united states: a cross-sectional study. *Public health*, 219:159–164, 2023.
- 17 J. Von Neumann and O. Morgenstern. *Theory of games and economic behavior*, 2nd rev. Princeton university press, 1947.
- 18 Wu Y. and et al. Incubation period of covid-19 caused by unique sars-cov-2 strains: A systematic review and meta-analysis. *JAMA*, 5(8), 2022.

How to Improve Joint Suitability Mapping for Search Space Reduction?

Haoyu Wang  

Department of Geography and the Environment, University of Texas at Austin, TX, USA

Jennifer A. Miller  

Department of Geography and the Environment, University of Texas at Austin, TX, USA

Abstract

Geoforensic analyses are used to identify the location history of objects or people of interest. An effective method for location history identification is to use joint probability or suitability of trace materials. Species distribution models have been used to derive joint suitability distributions using suitable biotic trace evidence such as pollen. One of the key objectives for such analyses is to effectively reduce potential search space and search effort for investigators. This research presents a novel framework for modeling the habitat suitability of pollen identified at the plant species-level to generate joint suitability maps. We provide major limitations and challenges faced by current geolocation analyses based on species distribution models, including opportunities to improve the joint suitability analyses for search space reduction. A conditional probability approach for geolocation identification is also demonstrated for possible future applications in real-world forensic cases.

2012 ACM Subject Classification Information systems → Geographic information systems

Keywords and phrases forensic geolocation, species distribution modeling, conditional probability, search space reduction

Digital Object Identifier 10.4230/LIPIcs.GIScience.2023.77

Category Short Paper

Funding The research was partly funded by DEVCOM ARL, ARO through a Multidisciplinary University Research Initiative Grant (#W911NF1910231). The research, interpretations, and perspectives reported here are those of the authors and should not be attributed to the Army or the Department of Defense.

1 Background

Environmental trace evidence helps link objects or people of forensic interest to time and locations [4]. One such useful candidate for trace evidence usually found on items at scenes is pollen because of their durability on multiple contact carriers such as soil, fabrics, and other materials [3]. The microbial and environmentally ubiquitous characteristics of pollen also make it easy to attach to surfaces. The ability to identify pollen is dramatically improved using DNA-based identification methods. For example, DNA metabarcoding with high-throughput sequencing technologies improved pollen identification in terms of both quantity and accuracy. This improvement can help generate high-resolution plant taxonomic results, leading to potentially more reliable applications using forensic evidence [2]. The practical use of biotic trace materials such as pollen and spores in forensic science has also been discussed in recent research [1, 2, 7]. New methods that estimate suitable habitats of pollen's parent plant taxa using species distribution models for geoforensic location analysis have also been introduced, but have not been widely used [9, 10]. Species distribution models are used in these studies to quantify species-environment correlations which can then be used



© Haoyu Wang and Jennifer A. Miller;
licensed under Creative Commons License CC-BY 4.0

12th International Conference on Geographic Information Science (GIScience 2023).

Editors: Roger Beecham, Jed A. Long, Dianna Smith, Qunshan Zhao, and Sarah Wise; Article No. 77; pp. 77:1–77:6

Leibniz International Proceedings in Informatics



LIPIC Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

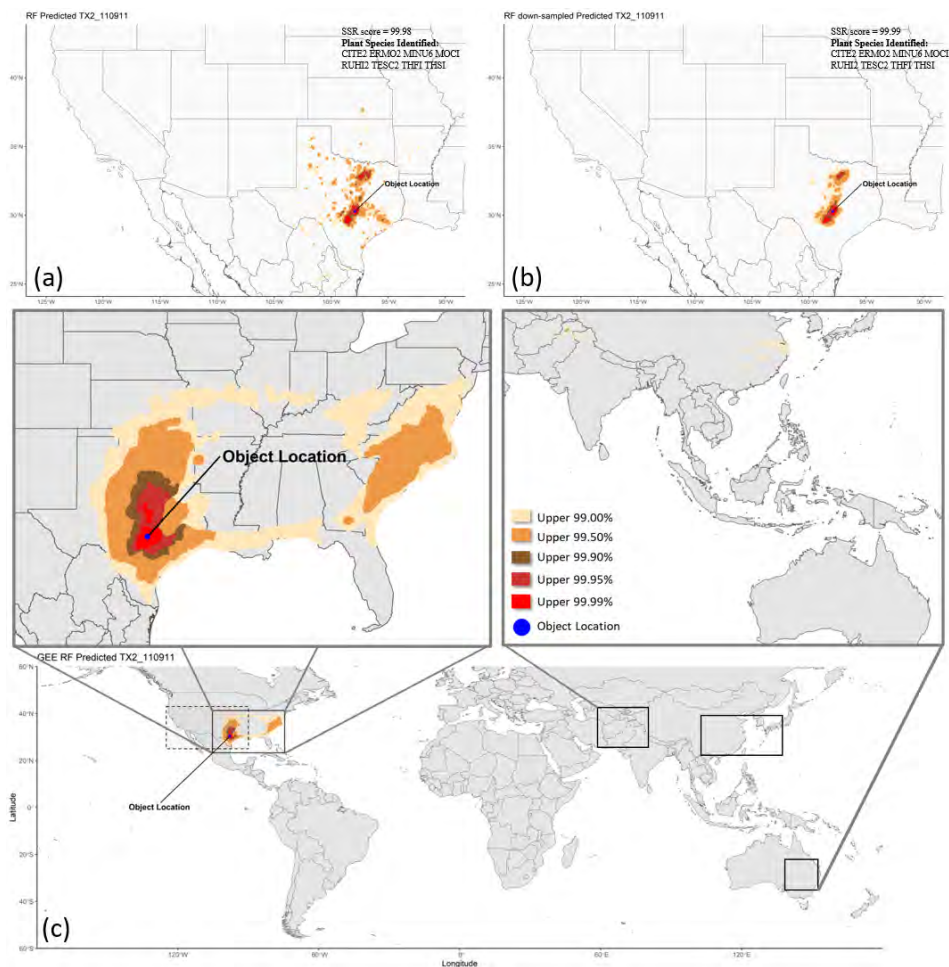
to predict the habitat suitability of plants and animals [5]. Joint suitability maps based on species distribution modeling results can then potentially reduce search areas and efforts for investigation purposes.

To test the feasibility of this joint suitability method, a study proposed a geographic attribution framework [10]. The authors collected bees in fieldwork and treated them as objects of interest, and the pollen grains sampled from the bees (pollen profiles of bees) were identified as trace evidence. Species distribution models were estimated for each identified species and combined to reduce the search space for an object that contained these species. Since the location of the bee (object) was known, the authors were able to assess the geolocation accuracy of models by quantifying and mapping the potential search space base on different percentiles. The authors used Google Earth Engine cloud-based geospatial platform that provides petabyte data and algorithms for fast computation to apply geographic attribution at a global scale. The inputs of this framework are georeferenced and filtered occurrences from the parent plant taxa of the recovered pollen species from the Global Biodiversity Information Facility, with more than 2.2 billion taxa information integrated from multiple data sources. The framework combines relative suitability distributions of taxa to a final prediction layer using a scaled-sum method, with percentiles indicating the priority of search areas for investigators, corresponding to different color hues as shown in Figure 1. These processes were also considered a set of methods for the *search space reduction* purpose. The *SSR score* in the top-right corner of Figure 1 shows the metric of joint suitability score, or search score, that indicates the performance of reducing the search space by comparing the joint suitability value between the object's location and all other locations. For more detailed explanations on the model building and accuracy assessment, see [10].

2 **Limitations**

Although the geographic attribution framework described here was useful when sufficient quantities of pollen are recovered from bee objects, some assumptions were made when we applied the search space reduction techniques, which bring limitations to the framework that was proposed in previous studies. The most noteworthy and challenging limitations of this framework are summarized below:

1. Current studies that use either probability- or suitability-based approaches (such as the use of species distribution models) to identify the geographic provenance of objects of interest have one common challenge. They can derive one best location or a series of probability-ranked locations. The top percentiles of location history such as the different percentiles/color hues of areas illustrated in Figure 1 are essentially a set of ranked search spaces. Study such as [8] has proposed methods to identify multiple traveled sites by objects of interest through solving geographic optimization problems, where suitability layers generated from species distribution models can be used as inputs. Although capturing any one portion of the total location history would be potentially helpful for investigation, discovering further methods to incorporate multiple location history identification is important. It is also hard for [10] to assess the location identification accuracy with information other than joint suitability of pollen, because the actual travel/foraging pattern or preference of each bee is hard to obtain.
2. Existing studies that generate the *joint* probability or suitability distributions of pollen's parent plants need to be retrospectively assessed for the distribution of each plant taxa. For geolocation analyses that involve combinations of multiple suitability layers, information may not be well analyzed through the combining process. For example,



■ **Figure 1** An example of joint suitability maps of the geographic location history identification of a bee object. The modeling results were made by two widely used species distribution models: (a) Random Forest, and (b) Random Forest down-sampled, at a subcontinental scale. This bee object has nine different pollen genus/species attached. (c): Joint suitability search areas at a global scale. The dashed box shows the subcontinental study area in (a) and (b). Solid boxes indicate potential search areas. Darker hues indicate areas with increasing joint suitability values.

although joint suitability of certain pollen profiles on an object of interest has returned high accuracy of geolocation identification results, additional steps are required to know which one pollen or group of pollen is contributing to the identification, or which pollen is adding noise to the identification.

3. For the geographic attribution framework tested in [10], the sampling locations of bees are assumed to be locations for accuracy assessments. However, a sampling location of a bee should be ideally treated as one of the location history stamps. Although this is not a problem in real-world applications since investigations would usually attempt to identify all meaningful location history of the objects of interest instead of focusing only on sampled/collected location, the misplaced *true* location could have an adverse impact on how we understand the geolocation analysis results.

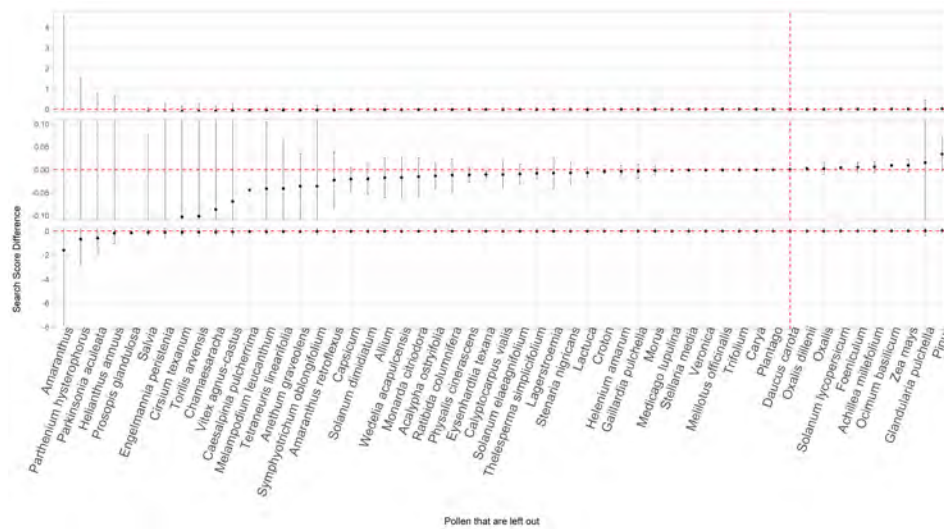
3 Updated Concepts

We provide two concepts based on the existing geographic attribution framework to potentially address some of the limitations mentioned above. For limitation #1, although the travel routes of bees are hard to obtain, this information could be partially available through inference or in some ways calculable in real-world forensic cases. Similar to [10], we set up a study area as a customized spatial domain, where i, j are longitude/latitude grid cells that have $M \cdot N$ total grid cells, where $i = 1, 2, \dots, M$ and $j = 1, 2, \dots, N$. For each grid location (i, j) , we use \mathcal{L} to denote the incident that people or objects of interest have traveled to this specific location. The conditional probability of people and objects that have traveled at a location (i, j) in a spatial domain is then provided as:

$$P(\mathcal{L}|T_1, T_2, \dots, T_n) = \frac{P(T_1, T_2, \dots, T_n|\mathcal{L}) \cdot P(\mathcal{L})}{P(T_1, T_2, \dots, T_n|\mathcal{L})P(\mathcal{L}) + P(T_1, T_2, \dots, T_n|\mathcal{L}^C)P(\mathcal{L}^C)} \quad (1)$$

where T_k is a set of the distribution of trace evidence such as the pollen or other biotic materials identified on objects of interest or at scenes, where $k = 1, 2, \dots, n$. Equation 1 is then illustrating how the pollen distribution probability provides an adjustment to probability surface derived by various investigation approaches, for example, criminal geographic targeting or geographic profiling that uses a set of locations from a series of crime [6]. The joint probability of equation 1 could be further computed with for example Bayesian inferences to solve the posterior probability which is the probability of people or objects have traveled to a location given that there is pollen found or corresponding plant taxa growing at this location. The minimal spatial unit for the calculation can be any meaningful size depending on the scales of focus, for example, a 900 m grid cell size used in the global geographic attribution cases.

To address the limitation #2 mentioned above, one would normally want to do repeated sampling of pollen profiles at one location, and need a method to distinguish and quantify the contribution of a single pollen within a pollen profile recovered on an object of interest. To achieve this, for every pollen profile of an object, one can keep one pollen out of the joint suitability combination and calculate the joint suitability score (search score) using the remaining pollen distribution layers, and repeat this process until every pollen found on this object is traversed. This is a methodology similar to leave-one-out cross validation, a procedure widely used in machine learning algorithms. To test the feasibility of this method, we first sampled multiple locations with various pollen profiles and fit species distribution models to obtain joint suitability search scores. Selected preliminary results from the leave-one-out method are shown in Figure 2. Each record at the x-axis of Figure 2 is the pollen that is left out in different pollen profiles. The mean search score difference on the y-axis is the difference in the two search space reduction scores before and after the corresponding pollen is left out. Negative score differences indicate that ignoring this pollen negatively affects geolocation identification, while positive score differences mean the opposite. We can then figure out how several pollen genus/species constantly contribute to or negatively affect the geolocation accuracy. For example, the genus of *Pinus* is always reducing the geolocation accuracy with a mean of around 0.03 for all geolocations we focused on. This corresponds to around five million pixels with a size of 900×900 m per pixel at a global scale. A possible reason for the negative contribution of *Pinus* is that pines as plants are growing in a large variety of environments and almost around the globe, contributing noise to most of the joint suitability analyses for geolocation identification. On the other hand, *Amaranthus* and other genus- and species-level pollen taxa with positive search contributions are having



■ **Figure 2** An example of search score differences after removing pollen from an object's pollen profile using joint suitability analyses. This example was made from species distribution modeling results computed from boosted regression trees (BRT) with multiple geolocations in a global spatial domain. The horizontal red dashed line indicates no search score changed from joint suitability analyses after ignoring this pollen taxa. Pollen taxa at the right side of the vertical red dashed line (including *Daucus carota*) indicate positive search score differences, while those at the left side have negative search score differences.

more fluctuated search score differences. This feature may suggest investigators carefully examine available information from different cases, including objects/people's possible ranges of activities, when such pollen taxa are present on objects or locations of interest.

References

- 1 Julia S. Allwood, Noah Fierer, and Robert R. Dunn. The Future of Environmental DNA in Forensic Science. *Applied and Environmental Microbiology*, 86(2):e01504–19, 2020. Publisher: American Society for Microbiology. doi:10.1128/AEM.01504-19.
- 2 Karen L. Bell, Kevin S. Burgess, Kazufusa C. Okamoto, Roman Aranda, and Berry J. Brosi. Review and future prospects for DNA barcoding methods in forensic palynology. *Forensic Science International. Genetics*, 21:110–116, March 2016. doi:10.1016/j.fsigen.2015.12.010.
- 3 Marzia Boi. Pollen attachment in common materials. *Aerobiologia*, 31(2):261–270, June 2015. doi:10.1007/s10453-014-9362-2.
- 4 D. C. Mildenhall. An unusual appearance of a common pollen type indicates the scene of the crime. *Forensic Science International*, 163(3):236–240, November 2006. doi:10.1016/j.forsciint.2005.11.029.
- 5 Jennifer A. Miller. Species distribution models: Spatial autocorrelation and non-stationarity. *Progress in Physical Geography: Earth and Environment*, 36(5):681–692, October 2012. Publisher: SAGE Publications Ltd. doi:10.1177/0309133312442522.
- 6 D. Kim Rossmo. *Geographic Profiling*. CRC Press, December 1999. Google-Books-ID: YQIS59Pv35oC.
- 7 Libby A. Stern, Jodi B. Webb, Debra A. Willard, Christopher E. Bernhardt, David A. Korejwo, Maureen C. Bottrell, Garrett B. McMahon, Nancy J. McMillan, Jared M. Schuetter, and Jack Hietpas. Geographic Attribution of Soils Using Probabilistic Modeling of GIS Data for Forensic Search Efforts. *Geochemistry, Geophysics, Geosystems*, 20(2):913–932,

77:6 How to Improve Joint Suitability Mapping for Search Space Reduction?

2019. __eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1029/2018GC007872>. doi:10.1029/2018GC007872.
- 8 Daoqin Tong, Tony H. Grubestic, Wangshu Mu, Jennifer A. Miller, Edward Helderop, Shalene Jha, Berry J. Brosi, and Elisa J. Bienenstock. Identifying the spatial footprint of pollen distributions using the Geoforensic Interdiction (GOFIND) model. *Computers, Environment and Urban Systems*, 87:101615, May 2021. doi:10.1016/j.compenvurbsys.2021.101615.
 - 9 Haoyu Wang, Jennifer A. Miller, Tony H. Grubestic, and Shalene Jha. A Framework for Using Ensemble Species Distribution Models for Geographic Attribution in Forensic Palynology. In *2022 IEEE International Symposium on Technologies for Homeland Security (HST)*, pages 1–7, November 2022. doi:10.1109/HST56032.2022.10025427.
 - 10 Haoyu Wang, Jennifer A. Miller, Tony H. Grubestic, and Shalene Jha. Using habitat suitability models for multiscale forensic geolocation analysis. *Transactions in GIS*, 27(3):777–796, 2023. __eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/tgis.13052>. doi:10.1111/tgis.13052.

Navigation in Complex Space: An Bayesian Nash Equilibrium-Informed Agent-Based Model

Yiyu Wang  

School of Geography, University of Leeds, UK

Jiaqi Ge  

School of Geography, University of Leeds, UK

Alexis Comber  

School of Geography, University of Leeds, UK

Abstract

This study proposed an improved pedestrian evacuation ABM employing Bayesian Nash Equilibrium (BNE) to simulate more realistic and representative individual evacuating behaviours in complex scenarios. A set of vertical blockades with adjustable gate widths was introduced to establish a simulation space with narrow corridor and bottlenecks and to evaluate the influences of BNE on individual navigation in complex space. To better match with the evacuating behaviours in real-world scenarios, the decision-making criterion of BNE evacuees was improved to a multi-strategy combination, with 80% of evacuees taking the optimal strategy, 15% taking sub-optimal strategy, and 5% taking the third-best one. The preliminary results demonstrate a positive impact of BNE on individual navigation in complex space, showing a distinct decrease of evacuation time with increasing proportion of BNE evacuees. The non-monotonicity of the variations in evacuation time also indicates the dynamic adaptability of BNE in addressing immediate challenges (i.e. blockades and congestions), which identifies alternative and potential faster paths during evacuations. A detailed description of the proposed ABM and an analysis of relevant experimental results are provided in this paper. Several limitations are also identified.

2012 ACM Subject Classification Computing methodologies → Intelligent agents; Computing methodologies → Modeling and simulation

Keywords and phrases Agent-based Modelling, Pedestrian Evacuation, Bayesian Nash Equilibrium, Individual Navigation, Complex Environment

Digital Object Identifier 10.4230/LIPICs.GIScience.2023.78

Category Short Paper

1 Introduction

Recent research has proposed a novel ABM for pedestrian evacuation which employs Bayesian Nash Equilibrium (BNE) to fill the gap of lacking representative and forward-looking individual behavioural models in relevant research on pedestrian evacuating simulations [5]. This ABM has been shown to be capable of producing more realistic individual evacuating behaviours in simple scenarios because evacuee agents following the BNE model are able to predict future congestion levels at each time step to find faster evacuation routes. The experimental results suggest that such model could better represent the real-world evacuating behaviours and improve the effectiveness of emergency management strategies.[5]

On this basis, this study aims to further evaluate the influences of the BNE model on individual navigation in complex spaces as well as its applicability in pedestrian simulations involving different scenarios. The above initial work has been extended by improving the decision-making logic of the initial BNE model in order to adapt to the evacuations in complex environments. An improved BNE-informed ABM has been developed in NetLogo with a series of vertical blockades with adjustable gate width brought into the simulation space to form complex evacuation scenarios. A series of simulation experiments were conducted to



© Yiyu Wang, Jiaqi Ge, and Alexis Comber;
licensed under Creative Commons License CC-BY 4.0

12th International Conference on Geographic Information Science (GIScience 2023).

Editors: Roger Beecham, Jed A. Long, Dianna Smith, Qunshan Zhao, and Sarah Wise; Article No. 78; pp. 78:1–78:6

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

examine how and whether the updated model plays roles in individual evacuation process in these complex spaces. The implementation details of this improved model, the analysis of the experimental results, as well as a discussion on limitations and further work have been provided in this paper.

2 Methodology

2.1 Theoretical Background

This research adopts a refined Bayesian Nash Equilibrium (BNE) [3] as the underlying theory of individual decision-making. BNE is a gaming strategy in which the players maximise their expected utilities and take the best strategy according to the probability distribution of other players' next decisions [1]. The BNE refinement takes incomplete information into account, which aligns with the situation in which some real-time information might be ignored by some pedestrians in real-world evacuation scenarios [3]. The proposed model embodies BNE as a set of utility functions and considers the probability distributions of neighbouring evacuees' further actions, to provide a more accurate representation of individual decision-making process in a complex evacuation scenarios, with potential applications in optimizing evacuation plannings in different scenarios.

2.2 Improved BNE Model

Bayesian Nash Equilibrium (BNE) was employed as a methodological framework to quantify the individual decision-making process in complex evacuation space. A series of utility functions have been incorporated into the improved ABM to implement the BNE behavioural model. Due to the non-sequential decision-making in BNE games [3], the evacuees following BNE model determined their future actions based on the values of Total Utility (U_{total}) for their neighbouring patches.

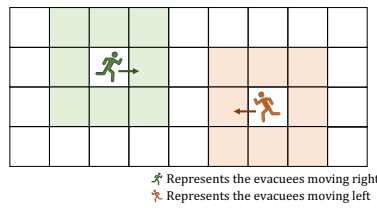
The concept of Total Utility (U_{total}) is associated with three crucial factors: Distance Utility (U_d), Comfort Utility (U_c), and Expected Comfort Utility (U_{ec}), and designated as the total value of U_d and U_{ec} , as represented by Eq. (1) [5]. This parameter, which value considers the distance to the exit, future congestion levels, and possible actions of other evacuees in their Moore neighbourhood¹, is participated in the decision-makings of BNE evacuees' next movements. That is, evacuees following BNE model are capable to avoid congested areas by forecasting possible movements of other nearby evacuees. For each BNE evacuee, the total utilities of all the passable patches in its Moore neighbourhood are calculated and compared to determine the favourable patch to move at the next time step (see Fig. 1). The BNE-related functions are described as follows.

$$U_{total} = U_d + U_{ec} \quad (1)$$

A. Distance Utility. U_d is associated with the distance between an evacuee's current position and the exit. It should be noted that the patches representing impassable barriers are not included in related calculations. The value is defined by a monotonically increasing function that approaches the maximum as evacuees are close to the exit point, as shown in Eq. (2).

$$U_d = (D - d)/D \quad (2)$$

¹ Moore Neighbourhood: a square-shaped neighbourhood with radius of one cell.



■ **Figure 1** The candidate patches of BNE evacuees.

Where, d represents the distance of the shortest route from the current location to the exit; D denotes the diagonal distance of the evacuation space.

B. Comfort Utility. U_c comprises a set of coefficients which is a fundamental component of U_{ec} . Its value is determined by the number of evacuees occupying the given patch and reflects the individual comfort level of this patch. An inverse relation can be found between the value of U_c and the crowd density of the certain patch, as illustrated in Eq. (3).

$$U_c = \begin{cases} 1.00, n \leq 2 \\ 0.51, n = 3 \\ 0.07, n = 4 \\ 0.00, n \geq 5 \end{cases} \quad (3)$$

Where, n represents the number of evacuees on the patch.

C. Expected Comfort Utility. U_{ec} is calculated as the product of two elements: Comfort utility U_c and the probability $p(n)$ that a particular number (n) of evacuees will move to the appointed patch at next time step, as shown in Eq. (4). The calculation of $p(n)$ considers not only the future movements of the evacuees on this patch, but also the possible actions of those located on the Moore neighbourhood.

$$U_{ec} = \sum_{n=0}^4 p(n)U_c(n) = \sum_{n=0}^4 C_N^n P_m^n (1 - P_m)^{N-n} U_c(n) \quad (4)$$

Where, n represents the count of evacuees on the patch at next time step; N denotes the total number of evacuees located on the patch and its Moore neighbourhood at this time step; P_m represents the probability of evacuees who may move to this patch at next time step, with a default value of 50%.

2.3 Improvement Details

In the initial implementation [5], the BNE model employed a decision-making criterion that required all the evacuees to choose the patch with maximum U_{total} , resulting in 100% of evacuees taking the best strategies. However, the experimental results indicated that this criterion could lead to all BNE evacuees located on the same patch making identical choices in the latter stages of simulations, which, in turn, resulted in localized congestion and reduced exiting speeds [5]. In this paper, this challenge is addressed by including some noise to the initial decision-making logic of BNE evacuees, by switching to a multi-strategy combination: with 80% of evacuees taking the optimal strategy (i.e. selecting the patch with highest U_{total}), 15% taking the sub-optimal strategy (i.e. choosing the patch with second-highest U_{total}), and 5% making the third-optimal choice (i.e. selecting the patch with third-highest U_{total}).

As well as the improved BNE model, two other behavioural models are included – Shortest Route (SR) model (Dijkstra’s search algorithm [2] was introduced to replace the weak SR strategy) and Random Follow (RF) model – as control groups in the proposed ABM to better evaluate the performances of BNE model [5].

3 Experimental Results

3.1 Model Initialisation and Implementation Details

The initial version of the BNE-informed ABM was developed in NetLogo and published at COMSES, which was retrieved from <https://doi.org/10.25937/75wf-aa82> [4]. The improved version is still in process and will be available once completed.

The initial configuration of the improved ABM involved the random dispersion of 2000 evacuees (agents) throughout the simulation space to the left side of the vertical blockades. The main purpose of this research is to explore the influences of different behavioural models on individual evacuating behaviours, with a specific focus on the capability of BNE evacuees to discover faster evacuation routes in order to navigate around congested areas and the barriers on their pathways. To achieve this objective, four distinct movement patterns were provided, including Shortest Route (SR), Random Follow (RF), BNE mixed with SR, and BNE mixed with RF. The percentage of BNE evacuees defaults to 100% in the last two BNE combinations, and the mixing ratios could be adjusted to meet the requirements of simulations.

To assess the effects of BNE on individual navigation in complex space, this paper conducted a simulation study where the individual evacuating behaviours could be observed in an evacuation space consisting of a narrow corridor defined by two vertical rectangular blockades with an adjustable-width gate for each barrier. All BNE-related utilities were computed at the beginning of each simulation and updated every time step. The decision-making criterion of BNE evacuees has been improved from a single strategy to a multi-strategy combination to better simulate the individual navigation in complex space with blockades, bottlenecks, and congestions.

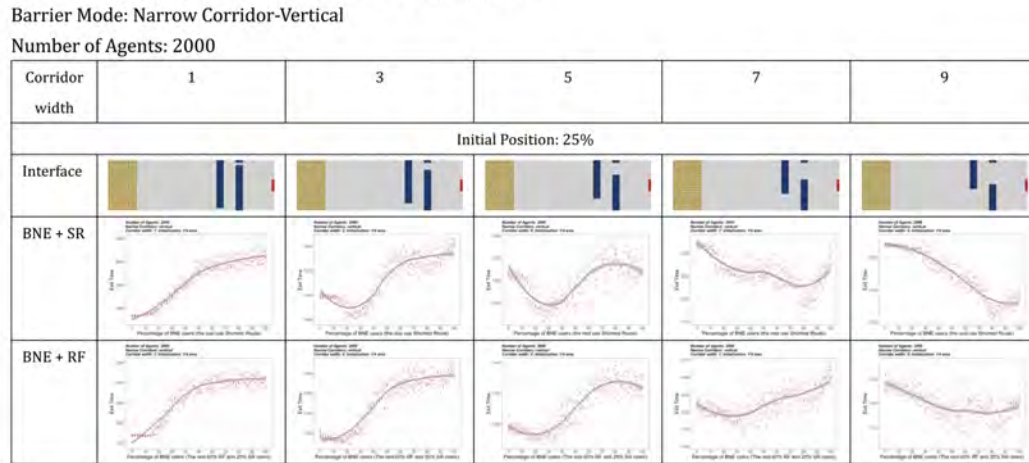
3.2 Simulation Experiments

A set of experiments simulated evacuations in a tunnel space consisting of vertical barriers with adjustable-sized gates. The model was initialized with 2000 evacuees, where the BNE evacuees were mixed with evacuees adhering to one of the other two behavioural models (i.e. SR and RF). The proportion of BNE agents was adjusted from 0% to 100% at 2% intervals, and the gate width for each blockade varied from 1 to 9 at 2-patch intervals. The simulations were replicated 10 times for each parameter configuration and stopped once all agents evacuate successfully through the exiting point. The exit time of each simulation was recorded to evaluate the impacts of BNE on individual evacuations in complex space.

3.3 Result Analysis

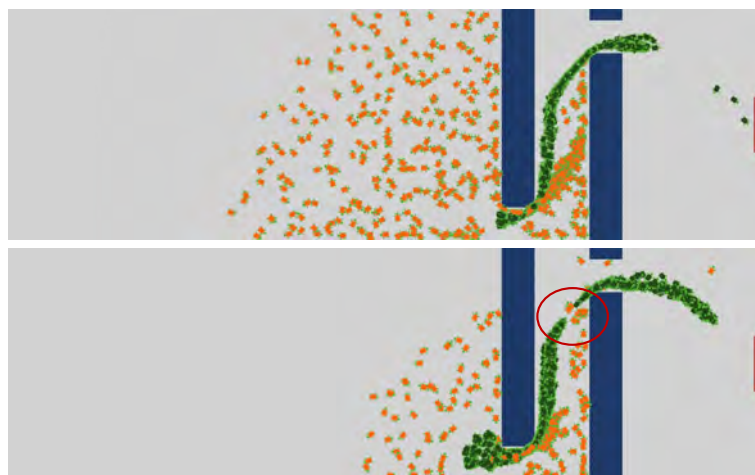
Fig.2 illustrates the variations in evacuation time of the evacuees following BNE with SR and with RF combinations respectively in a complex space with varied width of gates. A local line of fit with 95% confidence interval was also generated in the plots to reflect the relationship among exit time, percentage of BNE evacuees, and sizes of the barriers. As shown, there is little advantage of specifying BNE when the gate width of blockades is too narrow, while a decreasing trend of evacuation time with increasing proportion of BNE evacuees can be

observed in the scenarios with wider gates. That is, a positive impact of BNE on shortening evacuation time becomes salient as the increasing percentage of BNE agents participating into the simulations.



■ **Figure 2** Evacuation time against percentage of BNE-SR and BNE-RF combinations in complex scenarios with varied width of gates (2000 evacuees).

The non-monotonic changes of exit time against the two BNE combinations in complex scenarios are worthy of further discussion. A reasonable explanation for this phenomenon is that some of BNE evacuees may be trapped in the corner of the corridor as a queue of evacuees following other models was formed during evacuations. As shown in Fig.3, evacuees adhering to SR model (shown in green) were observed to follow an identical trajectory to avoid barriers and evacuate. The consistency of path selection may lead to a situation that BNE evacuees (shown in orange) were confined within the corridor as SR evacuees were stuck at the bottlenecks resulting in a heavy congestion.



■ **Figure 3** The model view of the evacuation process of BNE-SR combination.

However, this queue-induced deadlock is not permanent since the BNE agents were attempting to break free and navigate towards the relatively uncrowded area of the corridor, finding an alternative and potentially faster evacuation route. This also demonstrates the

dynamic adaptability of the improved BNE model in addressing the immediate challenges during evacuations (i.e. blockades, bottlenecks, and congestions) as well as discovering efficient evacuation pathways.

4 Discussion and Conclusion

This paper proposed an improved pedestrian evacuation model employing Bayesian Nash Equilibrium (BNE) within an ABM approach, with the objective of producing forward-looking and realistic individual evacuating behaviours in complex environments. The BNE-informed model integrates a set of vertical blockades with adjustable gate widths to establish a simulation space with narrow corridor and bottlenecks, providing a further evaluation of the influences of BNE on individual navigation in complex space. The decision-making criterion of BNE evacuees was improved to a multi-strategy combination, in which 80% of evacuees take the best strategy, 15% make the second-best decision, and 5% choose the third-best one, to improve the evacuation efficiency. The preliminary results indicate that BNE plays a positive role in individual navigation in complex scenarios involving bottlenecks and blockades, reflecting on the distinct decrease of evacuation time with the increasing proportion of BNE-guided evacuees. The non-monotonicity of the variations in exit time also reveals the dynamic adaptability of the BNE model in addressing immediate challenges such as barriers, bottlenecks, and congestions, as well as discovering efficient route during evacuations.

However, a few limitations still need to be addressed: 1) The non-monotonicity of the evacuation time revealed that introducing an appropriate proportion of BNE evacuees into simulations has a significant influence on reducing exit time, which need to be further studied and observed at the individual level; 2) Different types of barriers should be introduced to further assess the feasibility of the improved ABM in other complex scenarios; 3) Apart from exit time, the model needs to be evaluated in the terms of other parameters (e.g. comfort level, etc.) to provide a comprehensive evaluation of the role of BNE played in individual navigation under complex situations; 4) Real-world scenarios need to be introduced to examine the simulation accuracy of the improved ABM. The above issues will be gradually solved in the next step of this ongoing research.

References

- 1 S. Bouzat and M. N. Kuperman. Game theory in models of pedestrian room evacuation. *Phys. Rev. E*, 89:032806, March 2014. doi:10.1103/PhysRevE.89.032806.
- 2 DongKai Fan and Ping Shi. Improvement of dijkstra’s algorithm and its application in route planning. In *2010 seventh international conference on fuzzy systems and knowledge discovery*, volume 4, pages 1901–1904. IEEE, 2010.
- 3 Takashi Ui. Bayesian nash equilibrium and variational inequalities. *Journal of Mathematical Economics*, 63:139–146, 2016. doi:10.1016/j.jmateco.2016.02.004.
- 4 Yiyu Wang, Jiaqi Ge, and Alexis Comber. An agent-based simulation model of pedestrian evacuation based on bayesian nash equilibrium” (version 1.0.0). *CoMSES Computational Model Library*, 2022. doi:10.25937/75wf-aa82.
- 5 Yiyu Wang, Jiaqi Ge, and Alexis Comber. An agent-based simulation model of pedestrian evacuation based on bayesian nash equilibrium. *Journal of Artificial Societies and Social Simulation*, 26(3):6, 2023. doi:10.18564/jasss.5037.

Application of GIS in Public Health Practice: A Consortium's Approach to Tackling Travel Delays in Obstetric Emergencies in Urban Areas

Jia Wang¹

School of Computing & Mathematical Sciences,
University of Greenwich, London, UK

Peter M. Macharia

Department of Public Health,
Institute of Tropical Medicine, Antwerp, Belgium
Population & Health Impact Surveillance Group,
Kenya Medical Research Institute-Wellcome
Trust Research Programme, Nairobi, Kenya
Centre for Health Informatics, Computing, and
Statistics, Lancaster Medical School, Lancaster
University, UK

Kerry L. M. Wong

Faculty of Epidemiology and Population Health,
London School of Hygiene and Tropical Medicine,
UK

Uchenna Gwacham-Anisiobi

Nuffield Department of Population Health,
University of Oxford, UK

Abimbola Olaniran

Royal Tropical Institute, Amsterdam,
The Netherlands

Lenka Beňová

Department of Public Health,
Institute of Tropical Medicine, Antwerp, Belgium

Aduragbemi Banke-Thomas

Faculty of Epidemiology and Population Health,
London School of Hygiene and Tropical Medicine,
UK

Maternal and Reproductive Health Research
Collective, Lagos, Nigeria
School of Human Sciences,
University of Greenwich, London, UK

Itohan Osayande

School of Human Sciences,
University of Greenwich, London, UK

Prestige Tatenda Makanga

Surveying and Geomatics Department,
Faculty of Science and Technology,
Midlands State University, Gweru, Zimbabwe
Climate and Health Division, Centre for Sexual
Health and HIV/AIDS Research, Zimbabwe

Tope Olubodun

Department of Community Medicine and
Primary Care, Federal Medical Centre
Abeokuta, Abeokuta, Ogun, Nigeria

Olakunmi Ogunyemi

Lagos State Ministry of Health, Nigeria

Ibukun-Oluwa O. Abejirinde

Dalla Lana School of Public Health,
University of Toronto, Canada
Women's College Hospital Institute for
Health System Solutions and Virtual Care,
Toronto, Canada

Bosede B. Afolabi

Department of Obstetrics and Gynaecology,
College of Medicine, University of Lagos, Nigeria
Maternal and Reproductive Health Research
Collective, Lagos, Nigeria

Abstract

Geographic Information System (GIS) has become an effective and reliable tool for researchers, policymakers, and decision-makers to map health outcomes and inform targeted planning, evaluation, and monitoring. With the advent of big data-enabled GIS, researchers can now identify disparities

¹ Corresponding author



© Jia Wang, Itohan Osayande, Peter M. Macharia, Prestige Tatenda Makanga, Kerry L. M. Wong, Tope Olubodun, Uchenna Gwacham-Anisiobi, Olakunmi Ogunyemi, Abimbola Olaniran, Ibukun-Oluwa O. Abejirinde, Lenka Beňová, Bosede B. Afolabi, and Aduragbemi Banke-Thomas;
licensed under Creative Commons License CC-BY 4.0

12th International Conference on Geographic Information Science (GIScience 2023).

Editors: Roger Beecham, Jed A. Long, Dianna Smith, Qunshan Zhao, and Sarah Wise; Article No. 79; pp. 79:1–79:6



Leibniz International Proceedings in Informatics
LIPICIS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

and spatial inequalities in health at more granular levels, enabling them to provide more accurate and robust services and products for healthcare. This paper aims to showcase the progress of the On Tackling In-transit Delays for Mothers in Emergency (OnTIME) project, which is a unique collaborative effort between academia, policymakers, and industrial partners. The paper demonstrates how the limitations of traditional spatial accessibility models and data gaps have been overcome by combining GIS and big data to map the geographic accessibility and coverage of health facilities capable of providing emergency obstetric care (EmOC) in conurbations in Africa. The OnTIME project employs various GIS technologies and concepts, such as big spatial data, spatial databases, and public participation geographic information systems (PPGIS). We provide an overview of these concepts in relation to the OnTIME project to demonstrate the application of GIS in public health practice.

2012 ACM Subject Classification Information systems → Geographic information systems

Keywords and phrases GIS, Public Health, Accessibility, OnTIME, EmOC, Public Participation GIS, Big Data, Google

Digital Object Identifier 10.4230/LIPIcs.GIScience.2023.79

Category Short Paper

1 Introduction and Background

In 2020, the World Health Organisation estimated that around 287,000 women lost their lives worldwide due to maternal causes, which corresponds to nearly 800 maternal deaths occurring every day, or approximately one every two minutes [11]. This problem is particularly prevalent in Low- or Middle-Income Countries (LMICs), with Africa being particularly affected [11]. Africa not only accounts for the highest percentage of maternal deaths (69% of the global total of 287,000) and stillbirths (45% of the global total of 1.9 million), but it is also undergoing rapid urbanisation. Two-thirds of the world's population will live in urban areas by 2050 with a significant proportion of these additional 2.5 billion urban residents concentrating in Africa [9]. In urban settings, the odds of maternal death and stillbirth are significantly higher, partly explained by traffic-ridden journeys with longer travel time [5, 6, 4].

Timely access to emergency obstetric care provided by trained healthcare personnel can significantly reduce maternal deaths [12]. However, women with obstetric emergencies must travel to healthcare facilities capable of providing EmOC to access the needed care. Delays encountered during this journey from home to a healthcare facility providing EmOC have a significant impact on the health outcomes of both mothers and newborns [4, 10]. Many of these deaths are preventable with timely and effective intervention, highlighting the essentials of receiving EmOC in time. Therefore, there is a need to understand the travel time between the location where a need for obstetric emergency services arises and possible locations with EmOC facilities. This is particularly critical to prevent avoidable maternal deaths and stillbirths, as well as achieve the Sustainable Development Goals (SDG) for maternal and newborn mortality by 2030¹. However, the current approaches of estimating travel time, either reported or modelled estimates, do not accurately represent the dynamics of the journey between the women's location and EmOC facilities [7]. This is mainly due to lack of observational data on healthcare seeking behaviour to robustly parameterise the access

¹ <https://www.who.int/europe/about-us/our-work/sustainable-development-goals/targets-of-sustainable-development-goal-3>

models [3]. Thus, different dynamics such as traffic conditions, time of the day, weather variations, and other eventualities are not accounted for [3]. Further, majority of existing evaluations of travel time focus only on the public facilities, ignoring the significant role played by the private sector. In addition, these analyses evaluate travel time to the nearest facility, ignoring the possibility that a closer facility may be bypassed for alternative choices.

The OnTIME Consortium² is a cutting-edge partnership that brings together academics, decision-makers, and Google to offer solutions to the challenges encountered by pregnant mothers and caregivers in LMICs. The goal of the OnTIME project being delivered by the Consortium is to first assemble a geo-coded database of public and private hospitals with EmOC in major African conurbations and second, use this database to estimate close-to-reality travel times to the closest, second-closest, and third-closest facilities capable of EmOC services. A key deliverable of the OnTIME project is a digital dashboard³ (Figure 1) that enables policymakers to optimise the geographic accessibility of EmOC by providing more realistic estimations of travel time and geographic coverage within policy actionable units. The availability of more accurate coverage data represents the next frontier in policy-making and research for improving EmOC access in urban settings of LMICs [3]. The OnTIME project has a phased approach, starting with the most populated conurbations in Nigeria in phase 1. The second phase will focus on selected conurbations in Africa that have publicly available lists of health facilities with attributes of service provision, and eventually to other LMICs in Southeast Asia and Latin America (phase 3). In the completed first phase (2022-2023), focus group discussions and interviews were conducted with policymakers in Nigeria, and an online survey involving over 200 policymakers and researchers across Africa was carried out to obtain insights on the essential components and the implementation of the proposed dashboard. A facility functionality verification was conducted in 15 cities across public and private hospitals in Nigeria, which were selected based on a population of at least 1 million in 2022 or projected to reach 1 million by 2030. This effort led to the development of the digital dashboard that displays the time it takes for pregnant women to access EmOC of different levels in the selected urban areas of Nigeria. The displayed travel catchment areas reflect the functional geographical coverage and accessibility of EmOC, indicating areas of inequitable access that require prioritisation. This dashboard will inform and catalyse policy actions to improve geographical accessibility, contribute to Nigeria's commitment to universal health coverage and SDG 3, and ultimately lead to reduced maternal and perinatal mortality. There is an open database of generated travel time accompanying the dashboard (discussed in 2.1). The subsequent sections illustrate the GIS approaches employed in the initial phase of the OnTIME project.

2 Spatial Mapping and Big Data

Spatial mapping is commonly employed in public health to offer valuable understandings into the arrangement and availability of healthcare resources. This data can guide choices concerning the positioning of healthcare facilities and the distribution of healthcare resources. One of the current advancements in GIS is coming from our increasing capability of collecting, storing, processing and visualising mass volume of information of great complexity, a phenomenon known as “big data” [8]. By including the geographic coordinates in big data, it becomes big spatial data. As stated by the UK's national mapping agency Ordnance Survey, “with this additional spatial dimension, much deeper insights about the records in a dataset, and their relationships can often be drawn” [13].

² <https://www.ontimeconsortium.org/>

³ <https://emergencyobstetriccare.webapps.google.com/overview>



■ **Figure 1** OnTIME project's core deliverable: a digital dashboard displaying geographic access to emergency obstetric care designed for policymakers.

The OnTIME project involves several big spatial datasets. The data retrieved from Google Directions API includes spatial data on road networks, real-time and historical traffic data. Our industrial partner, Google, incorporated this vast amount of data to generate realistic routes between locations and predict the driving time between settlements and facilities providing EmOC. From the retrieved data, the following summaries can be computed: i) travel times to the first, second, and third nearest EmOC facilities, ii) the count of health facilities within 15, 30, and 60 minutes of driving time, and iii) proportion of women of child bearing age (15-49 years) within the same time thresholds. These time estimates are disaggregated by facility ownership (public, private and a combination of both), for eight traffic scenarios at different time of day and day of week.

The project also includes a large geo-coded dataset⁴ that stores information on the functionality and capability of both public and private hospitals offering EmOC from 15 cities in Nigeria. In the second and third phases of the project, comparable large-scale spatial databases will be constructed, concentrating on other LMIC urban areas. The extent of the 15 Nigerian cities was established by using spatial overlays to cross-reference the shapefiles of the local government area (LGA) boundaries⁵ with WorldPop⁶'s gridded surface of population (at a resolution of 100 m²), Google Maps, and the Global Human Settlement⁷ layers showing gridded surfaces of urban areas. Where applicable, locals were consulted to confirm the results. The centroids (as origins) of a 600 m² gridded dataset covering the entire study region were used to compute routes to the nearest EmOC facilities (as destination) using the Google Directions API.

3 Public Participation GIS

PPGIS is a “field within geographic information science that focuses on ways the public uses various forms of geospatial technologies to participate in public processes, such as mapping and decision making” [14]. Surveys and interviews are frequently used tools in PPGIS as a way to engage with stakeholders and gather their opinions and feedback on particular

⁴ <https://www.ontimeconsortium.org/relevant-databases>

⁵ <https://grid3.org>

⁶ <https://www.worldpop.org/>

⁷ <https://ghsl.jrc.ec.europa.eu/>

issues or topics. To ensure a fit-for-purpose dashboard, we conducted an online survey with policymakers and researchers to understand key considerations needed for developing a policy-ready dashboard of geospatial access to EmOC in Africa. We gathered information about participants' knowledge of the locations where poor geographic accessibility to EmOC is a concern, the technological resources currently utilised for EmOC service planning, their dashboard feature preferences, and the possibility of a dashboard to tackle the issue of inadequate EmOC accessibility.

Moreover, government stakeholders at both the state and federal levels in Nigeria were involved through a combination of in-person and virtual semi-structured interviews. These interviews were conducted with six policymakers and 17 senior civil servants, representing seven states across five geopolitical zones and the Federal Capital Territory, Abuja. Results [2] suggests that technocrats recognise the ideal of data-driven needs assessment in enhancing maternal care, although this is frequently impacted by various factors such as political pressures, persistent community advocacy, irregular short-term administrative cycles, and donor-driven funding decisions. Despite the possibility of obstacles, there is substantial enthusiasm and acceptance for the use of GIS-enabled dashboard to aid in health planning, particularly in circumstances where innovation and technology are already ingrained in the current government's administrative approach [2].

4 Discussion and Future Work

This paper showcases how GIS can be used to collect, represent, and reason data, leading to better public health planning and decision-making. Specifically, it illustrates the use of spatial mapping, big data, spatial database and PPGIS to address maternal care delays in the OnTIME project. To enable more informed decision-making, it is essential to incorporate geospatial data and leverage advanced GIS techniques. The potential future work includes:

- A further in-depth investigation of access inequality between regions can be conducted. Spatial auto-correlation could be used to assess the spatial agglomeration characteristics of accessibility. To explore the underlying reasons for the imbalance in geographic accessibility, socio-economic factors are extracted, thus enabling further analyses to investigate spatial associations by spatial statistical techniques such as geographically weighted regression.
- The adoption of the spatiotemporal exploratory data analysis in the project. This is a methodological approach to detect and describe patterns, trends, and relations in data in both space and time [1]. Spatiotemporal analysis can be used to examine the trends and patterns of EmOC utilisation and need over time.
- Propose an ontology that identifies and decomposes geographic access elements of maternal healthcare into a hierarchy of categories, which is further systematised using extensions of existing formal ontologies. This way, we can provide a methodology- and context-independent measure of geographic accessibility that could then be used to extrapolate conceptual models for a variety of wider public health applications.

Ultimately, the OnTIME Consortium is committed to contribute to global efforts to reduce maternal mortality by generating closer-to-reality assessments of geographic access gaps to critical maternal health services. The Phase I has already strongly shown that a collaborative and participatory approach makes GIS data more meaningful and yields greater impact, especially in a time in which the global community is committed to “leave no one behind”.


References

- 1 Natalia V. Andrienko and Gennady L. Andrienko. *Exploratory analysis of spatial and temporal data - a systematic approach*. Springer, 2006. doi:10.1007/3-540-31190-4.
- 2 Aduragbemi Banke-Thomas, Ibukun-Oluwa Omolade Abejirinde, Olakunmi Ogunyemi, and Uchenna Gwacham-Anisiobi. Innovative dashboard for optimising emergency obstetric care geographical accessibility in nigeria: Qualitative study with technocrats. *Health Policy and Technology*, 12(2):100756, 2023. doi:10.1016/j.hlpt.2023.100756.
- 3 Aduragbemi Banke-Thomas, Peter M. Macharia, Prestige Tatenda Makanga, Lenka Beňová, Kerry L. M. Wong, Uchenna Gwacham-Anisiobi, Jia Wang, Tope Olubodun, Olakunmi Ogunyemi, Bosede B. Afolabi, Basseyy Ebenso, and Ibukun-Oluwa Omolade Abejirinde. Leveraging big data for improving the estimation of close to reality travel time to obstetric emergency services in urban low- and middle-income settings. *Frontiers in Public Health*, 10, 2022. doi:10.3389/fpubh.2022.931401.
- 4 Aduragbemi Banke-Thomas, Cephas Ke on Avoka, Uchenna Gwacham-Anisiobi, and Lenka Benova. Influence of travel time and distance to the hospital of care on stillbirths: a retrospective facility-based cross-sectional study in lagos, nigeria. *BMJ Global Health*, 6(10), 2021. doi:10.1136/bmjgh-2021-007052.
- 5 Aduragbemi Banke-Thomas, Cephas Ke on Avoka, Uchenna Gwacham-Anisiobi, Olufemi Omololu, Mobolanle Balogun, Kikelomo Wright, Tolulope Temitayo Fasesin, Adedotun Olusi, Bosede Bukola Afolabi, and Charles Ameh. Travel of pregnant women in emergency situations to hospital and maternal mortality in lagos, nigeria: a retrospective cohort study. *BMJ Global Health*, 7(4), 2022. doi:10.1136/bmjgh-2022-008604.
- 6 Aduragbemi Banke-Thomas, Kerry L M Wong, Francis Ifeanyi Ayomoh, Rokibat Olabisi Giwa-Ayedun, and Lenka Benova. “in cities, it’s not far, but it takes long”: comparing estimated and replicated travel times to reach life-saving obstetric care in lagos, nigeria. *BMJ Global Health*, 6(1), 2021. doi:10.1136/bmjgh-2020-004318.
- 7 Aduragbemi Banke-Thomas, Kerry L M Wong, Lindsey Collins, Abimbola Olaniran, Mobolanle Balogun, Ololade Wright, Opeyemi Babajide, Babatunde Ajayi, Bosede Bukola Afolabi, Akin Abayomi, and Lenka Benova. An assessment of geographical access and factors influencing travel time to emergency obstetric care in the urban state of Lagos, Nigeria. *Health Policy and Planning*, 36(9):1384–1396, August 2021. doi:10.1093/heapol/czab099.
- 8 Michael F Goodchild. *GIS in the Era of Big Data*. CNRS-UMR Géographie-cités 8504, 2016. URL: <http://journals.openedition.org/cybergeo/27647>.
- 9 United Nations et al. World urbanization prospects: The 2018 revision. [Online; posted 2019].
- 10 Friday Okonofua, Donald Imosemi, Brian Igboin, Adegboyega Adeyemi, Chioma Chibuko, Adewale Idowu, and Wilson Imongan. Maternal death review and outcomes: An assessment in lagos state, nigeria. *PloS one*, 12(12):e0188392, 2017. doi:10.1371/journal.pone.0188392.
- 11 World Health Organization et al. Trends in maternal mortality 2000 to 2020: estimates by who, unicef, unfpa, world bank group and undesa/population division, 2023. [Online; posted 23-03-2023].
- 12 Anne Paxton, Deborah Maine, Lynn Freedman, Deborah Fry, and Samantha Lobis. The evidence for emergency obstetric care. *International Journal of Gynecology & Obstetrics*, 88(2):181–193, 2005. doi:10.1016/j.ijgo.2004.11.026.
- 13 Ordnance Survey. What is spatial data?, April 2023. URL: <https://www.ordnancesurvey.co.uk/blog/what-is-spatial-data> [cited April 20 2023].
- 14 David Tulloch. Public participation gis (ppgis). In *Encyclopedia of Geographic Information Science*. SAGE Publications, 2008.

The Ups and Downs of London High Streets Throughout COVID-19 Pandemic: Insights from Footfall-Based Clustering Analysis

Xinglei Wang¹ ✉ 

SpaceTimeLab for Big Data Analytics, University College London, UK

Xianghui Zhang ✉ 

SpaceTimeLab for Big Data Analytics, University College London, UK

Tao Cheng ✉ 

SpaceTimeLab for Big Data Analytics, University College London, UK

Abstract

As an important part of the economic and social fabric of urban areas, high streets were hit hard during the COVID-19 pandemic, resulting in massive closures of shops and plunge of footfall. To better understand how high streets respond to and recover from the pandemic, this paper examines the performance of London's high streets, focusing on footfall-based clustering analysis. Applying time series clustering to longitudinal footfall data derived from a mobile phone GPS dataset spanning over two years, we identify distinct groups of high streets with similar footfall change patterns. By analysing the resulting clusters' footfall dynamics, composition and geographic distribution, we uncover the diverse responses of different high streets to the pandemic disruption. Furthermore, we explore the factors driving specific footfall change patterns by examining the number of local and nonlocal visitors. This research addresses gaps in the existing literature by presenting a holistic view of high street responses throughout the pandemic and providing in-depth analysis of footfall change patterns and underlying causes. The implications and insights can inform strategies for the revitalisation and redevelopment of high streets in the post-pandemic era.

2012 ACM Subject Classification Applied computing → Sociology

Keywords and phrases High street, performance, footfall, clustering analysis, COVID-19

Digital Object Identifier 10.4230/LIPICs.GIScience.2023.80

Category Short Paper

Funding *Xinglei Wang*: the author's PhD research is jointly funded by China Scholarship Council and the Dean's Prize from University College London.

1 Introduction

Over the past 3 years, the COVID-19 pandemic, a global public health crisis of unprecedented scale, has brought about substantial changes to urban environment [7]. Although the COVID-19 is no longer a public health emergency of international concern [9] by the time this paper was written, the long-term effects linger and continue to shape urban landscapes.

Among the most affected urban areas are spaces of consumption such as high streets which are often the heart of local communities, serving as centres for commerce, social interaction, and cultural activities. Important as they are, high streets across the UK suffered a devastating blow during the pandemic, with over 17500 chain stores and other venues closing in Great Britain [2] and footfall decreasing by over 80% [5]. Given the pivotal role of

¹ Corresponding author



high streets in the economic and social fabric of urban systems, it is of great importance to explore the impacts of pandemic on high streets' performance to comprehend the broader consequences on local economies, businesses, and communities.

Existing studies have utilised various forms of data to assess the economic performance of consumption spaces, such as vacancies [3] and footfall [8]. Within the COVID-19 context, Enoch et al. used footfall data to analyse the impact of COVID-19 pandemic on six high streets in England [4]. Ballantyne et al. used a mobile phone app location dataset to examine the recent recovery of retail centres from the pandemic [1]. Although they provided empirical evidence of the impact of the pandemic, there are some limitations remained to be addressed. Firstly, the COVID-19 pandemic and its aftereffects last several years, with several rounds of national lockdowns enacted, but existing studies do not cover the whole period, thus cannot present the full picture of the responses of high streets. Secondly, existing literature only focus on the change of footfall counts, lacking in-depth analysis of the change patterns and its underlying causes. To fill the research gaps, we examine and evaluate the performance of London's high streets during the pandemic using longitudinal footfall data. Specifically, footfall data spanning over two years was calculated from a mobile phone GPS dataset and time series clustering was applied to generate multiple groups of high streets with similar patterns. By analysing the distinctive footfall change patterns of resulting clusters and their geographic distribution, we unravelled the varying responses of different high streets to the disruption and the spatial patterns of different clusters of high streets. Furthermore, we linked the clustering results to the existing typology of retail centres and gained further insight into how the different composition of clusters corresponds to their performance. Lastly, we delved into the cause of particular change patterns by looking at the number of local and nonlocal visitors.

In the following sections, we describe the dataset used in this study, followed by a brief introduction of the methods employed. We present and analyse the results, discuss their implications, and provide recommendations for policymakers and urban practitioners. We conclude the paper by summarising the main findings and pointing out directions for future work.

2 Data

- **Mobile phone GPS trajectory data:** it is a large-scale mobility dataset which contains millions of anonymous users' mobile phone GPS trajectory data (collected from tens of location-based service apps) provided by Location Sciences under GDPR compliance. The dataset spans 3 years, and we define our study period from the first Monday of February 2020 (03/02/2020) to the last Sunday of April 2022 (24/04/2022), spanning 812 days (116 weeks). The number of unique devices in London in February 2020 exceeds 610,000. The data collection method and sampling rate over the whole country remains consistent throughout the study period.
- **High street boundary dataset:** provided by the Greater London Authority², this is a shapefile containing the boundaries of 616 London high streets located outside the Central Activity Zone.
- **Lower Layer Super Output Areas (LSOAs):** It is a geographic hierarchy designed to improve the reporting of small area statistics in England and Wales. This study utilised the LOSAs dataset created in the most recent 2021 census and only those within Greater London area were included.

² <https://data.london.gov.uk/dataset/gla-high-street-boundaries>

3 Methods

In this section, we present the workflow of footfall calculation and give a brief introduction to the K-means clustering method.

Footfall calculation

The workflow consists of the following steps:

1. Home detection: obtain the LSOA-level home location of each individual, which is denoted as *home_lsoa*. Here, the home of a person is defined as the LSOA where they generate the greatest number of GPS points during night time (e.g., 22:00-07:00).
2. Stop detection: to get the stop which is where people remain stationary for more than a specific amount of time (we set 5 minutes as the threshold in this study).
3. Identity inference: infer the identities (being one of resident and non-resident) of the people visiting a certain high street. If the *home_lsoa* of a person is one of the LSOAs that intersects with the high street, then this person is considered as a local resident, otherwise, a non-resident.
4. Footfall calculation: join people's stops with high street boundaries and calculate the footfall w.r.t resident and non-resident by day and sum the 2 types to get the overall footfall.

The output is daily footfall counts on each of the high street in London. We further aggregate the daily footfall into weekly ones by summing over 7 days of each week. This step can not only smooth the time series but also reduce the length of it, which can improve the results of clustering. We also normalise the footfall counts to convert them into relative values between 0 and 1, which is an essential pre-processing step for clustering.

K-means time series clustering

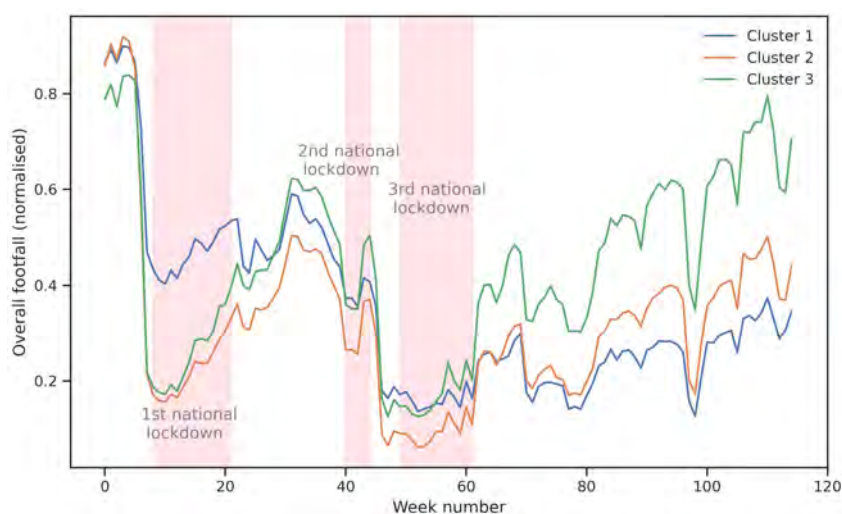
We employ a K-means time-series clustering algorithm to cluster the time-series of weekly overall footfall on the high streets. We use the Euclidean distance as the metric for clustering. One advantage of this method is that the distance between any two objects is not affected by the addition of new objects to the analysis, which may be outliers [6]. Elbow method and the Silhouette score are used to identify the optimal cluster number K.

4 Results

High street clusters and their spatial distribution and composition

Three high street clusters are identified based on the time-series pattern in footfall. The share of high streets in each cluster are 39 %, 33 % and 28 %, respectively. Figure 1 shows the footfall time series for the entire study period across the three clusters. Cluster 1 had the smallest drop when the first national lockdown came into effect. But after the third national lockdown, it remained the lowest while the other two (especially Cluster 3) recovered better. Cluster 2 and 3 exhibited similar trend, but Cluster 3 surpassed Cluster 2 (and Cluster 1) significantly during the “stepping out of lockdown” period, reaching its pre-pandemic level. The results demonstrate the varied ability of high streets to weather crises.

Figure 2 shows the spatial distribution of the clusters where some degree of spatial clustering is notable. Most of the high streets in Cluster 1 are located in inner London area, while Cluster 3 finds more high streets in outer London. Combining the spatial distribution



■ **Figure 1** Time-series pattern of the three identified high street clusters (the pink shades indicate the period of three national lockdowns).

and performance, We can draw the crude conclusion that high streets located at the periphery of the city tend to recover better than those closer to the city centre. High streets in Cluster 2 are more evenly distributed, but are relatively bigger in sizes.

To gain more information about what types of high streets each cluster contains, we further look into the composition of each cluster by linking the high streets with Retail Centre Typology provided by CDRC³. Figure 3 shows the composition of each cluster, where we can see that Cluster 1 has the highest proportion of small local centres compared to Cluster 2 and 3. In Cluster 2, only 38.7% of the high streets are small local centres, the lowest among the three clusters, while higher percentages of district centre and town centre are found in this cluster. As for Cluster 3, the composition is very similar to Cluster 1, but the proportion of high streets located in the outer London area is much higher than that of Cluster 1 (referring to Figure 2).

Local and nonlocal visitors

It is of great interest to us to uncover the underlying cause that made the three clusters affected so disproportionately by the pandemic. In particular, the compositions of Cluster 1 and 3 are very similar, yet they have such distinctive responses during multiple rounds of lockdowns and after-lockdown recovery period. With the question in mind, we calculated the number of local visitors (residents) and nonlocal visitors (non-residents) in each cluster and present the result in Figure 4.

Clearly, the stronger resilience Cluster 1 showed during the first national lockdown is owing to the preservation of local residents, while its downfall in the recovery phase is largely due to the continued loss of local residents (possibly because of people moving out of city [10]). The rise of Cluster 3 after the third national lockdown is much explained by the rapid increase in the number of both local residents and nonlocal visitors.

³ <https://data.cdrc.ac.uk/dataset/retail-centre-boundaries-and-open-indicators>

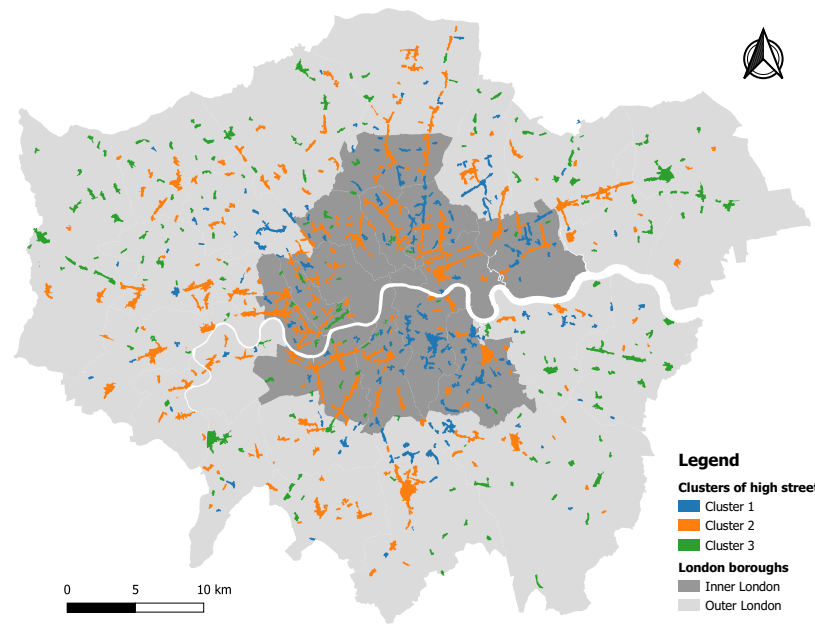


Figure 2 Spatial distribution of high streets in three clusters.

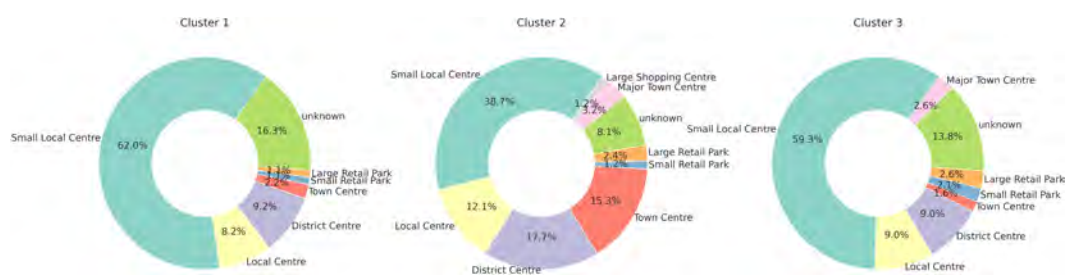


Figure 3 The proportion of different high streets in three clusters.

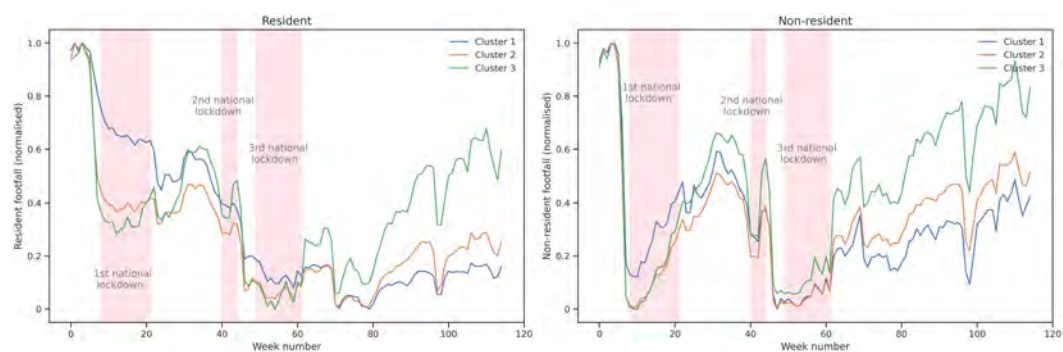


Figure 4 The local (resident) and nonlocal (non-resident) visitors in three clusters.

5 Discussion and conclusions

By analysing the full trajectory of high street footfall, we made a significant discovery that the resilience of high streets towards pandemic is very complicated both in space and time: in general, high streets located at the periphery of the city have recovered better and those which endured better through lockdowns may not recover well in post-pandemic period. We also made the first attempt to uncover the underlying cause of such varied responses of high streets. The interesting finding that local residents is the “key to success” highlights the importance of community engagement for London high streets. Policymakers and local authorities might consider organising more local events and activities, as well as launching initiatives such as community-led regeneration projects, to help strengthen the high streets’ attractiveness, bring back residents and create a sense of ownership and pride among them.

In conclusion, this paper is a first step towards the quantification and clustering analysis of high streets performance throughout the COVID-19 pandemic. By identifying the variations in footfall among different high streets, it provides evidence-based insights for decision-making processes related to urban regeneration, infrastructure development, and the formulation of policies that support local businesses. Policymakers can tailor interventions and allocate resources more effectively, ensuring a targeted approach that addresses the unique characteristics and needs of each high street. For future work, we will incorporate more contextual features into our clustering analysis, such as catchment demographics and built environment information to investigate the mechanisms in which high street response to disruptions.

References

- 1 Patrick Ballantyne, Alex Singleton, and Les Dolega. Using unstable data from mobile phone applications to examine recent trajectories of retail centre recovery. *Urban informatics*, 1(1):21, 2022.
- 2 BBC. Pandemic impact ‘yet to be felt’ on high streets. <https://www.bbc.co.uk/news/business-56378667>, 2021. Accessed: 2023-05-22.
- 3 Les Dolega and Alex Lord. Exploring the geography of retail success and decline: A case study of the liverpool city region. *Cities*, 96:102456, 2020.
- 4 Marcus Enoch, Fredrik Monsuur, Garyfalia Palaiologou, Mohammed A Quddus, Fiona Ellis-Chadwick, Craig Morton, and Rod Rayner. When covid-19 came to town: Measuring the impact of the coronavirus pandemic on footfall on six high streets in england. *Environment and Planning B: Urban Analytics and City Science*, 49(3):1091–1111, 2022.
- 5 Retail Gazette. Uk footfall drops to lowest level on record. <https://www.retailgazette.co.uk/blog/2020/04/uk-footfall-drops-to-lowest-level-on-record/>, 2020. Accessed: 2023-05-22.
- 6 Ali Javed, Byung Suk Lee, and Donna M Rizzo. A benchmark study on time series clustering. *Machine Learning with Applications*, 1:100001, 2020.
- 7 Ka Yan Lai, Chris Webster, Sarika Kumari, and Chinmoy Sarkar. The nature of cities and the covid-19 pandemic. *Current Opinion in Environmental Sustainability*, 46:27–31, 2020.
- 8 Christine Mumford, Cathy Parker, Nikolaos Ntounis, and Ed Dargan. Footfall signatures and volumes: Towards a classification of uk centres. *Environment and Planning B: Urban Analytics and City Science*, 48(6):1495–1510, 2021.
- 9 WHO. Who chief declares end to covid-19 as a global health emergency. <https://news.un.org/en/story/2023/05/1136367>, 2023. Accessed: 2023-05-22.
- 10 Elias Willberg, Olle Järv, Tuomas Väisänen, and Tuuli Toivonen. Escaping from cities during the covid-19 crisis: Using mobile phone data to trace mobility in finland. *ISPRS international journal of geo-information*, 10(2):103, 2021.

Agent-Based Modeling of Consumer Choice by Utilizing Crowdsourced Data and Deep Learning

Boyu Wang¹ ✉ 🏠 

Department of Geography, University at Buffalo, NY, USA

Andrew Crooks ✉ 🏠 

Department of Geography, University at Buffalo, NY, USA

Abstract

People's opinions are one of the defining factors that turn spaces into meaningful places. Online platforms such as Yelp allow users to publish their reviews on businesses. To understand reviewers' opinion formation processes and the emergent patterns of published opinions, we utilize natural language processing (NLP) techniques especially that of aspect-based sentiment analysis methods (a deep learning approach) on a geographically explicit Yelp dataset to extract and categorize reviewers' opinion aspects on places within urban areas. Such data is then used as a basis to inform an agent-based model, where consumers' (i.e., agents') choices are based on their characteristics and preferences. The results show the emergent patterns of reviewers' opinions and the influence of these opinions on others. As such this work demonstrates how using deep learning techniques on geospatial data can help advance our understanding of place and cities more generally.

2012 ACM Subject Classification Computing methodologies → Natural language processing; Computing methodologies → Agent / discrete models; Social and professional topics → Geographic characteristics

Keywords and phrases aspect-category sentiment analysis, consumer choice, agent-based modeling, online restaurant reviews

Digital Object Identifier 10.4230/LIPIcs.GIScience.2023.81

Category Short Paper

Supplementary Material *Software (Source Code)*: <https://github.com/wang-boyu/yelp-abm>
archived at `swh:1:dir:e6398b2a2185fab1b5168bfd588d75261bac2df1`

1 Introduction

People's opinions about places reflect their emotional attachment to locations that hold meanings to them. With the rise of social media platforms such as Google Reviews, TripAdvisor, and Yelp, vast numbers of opinions about local businesses, including restaurants, have been published online. These text reviews provide valuable insights into various aspects of the dining experience, such as the quality of food and service. Studying these reviews through aspect-based sentiment analysis (ABSA) can help identify key aspects that customers care about and understand the rationales behind customers' needs and preferences. The present work aims to address the following research questions (RQ): How to utilize recent advancements in natural language processing (NLP) techniques to: 1) help identify key aspects that customers care about when choosing restaurants, and 2) help inform consumer choice modeling in the context of visitation patterns to restaurants?

Over the last several decades, a body of literature has grown with respect to studying consumers' choice factors when choosing which restaurant to visit. For example, Auty [2] identified several main choice variables (e.g., food type, food quality, location, etc.) in

¹ corresponding author

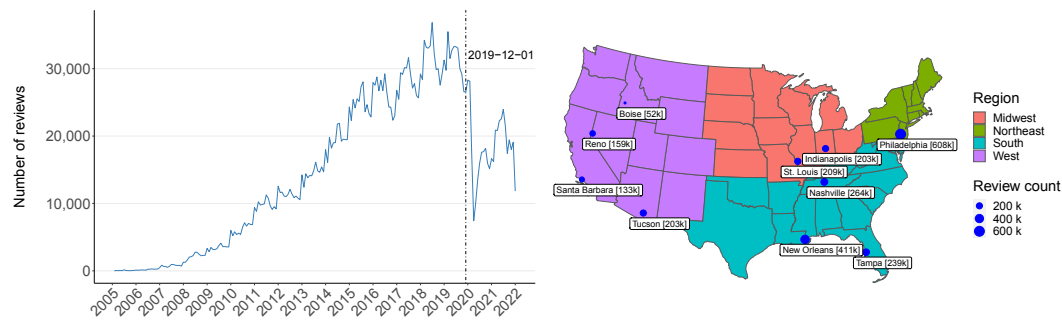


the restaurant selection processes, and how their relative importance varied with respect to consumers' demographic segments such as age and income. Focusing on quick-casual restaurants, Ryu and Han [10] analyzed the relationships between restaurant qualities (i.e., food, service, and physical environment) and consumers' perceived price, and how they in turn affected customer satisfaction and their behavioral intention. Similarly, Bujisic et al. [3] investigated the interactions between restaurant qualities and their effects on consumers' intentions. However, most of the existing studies have used qualitative methods (e.g., surveys, interviews, focus groups) to collect consumer responses. As a result, they are limited by small sample sizes of a few hundred people, and the scope of these studies are either towards specific demographic groups (e.g., students, senior citizens) or specific types of restaurants (e.g., fast-food, Chinese, Korean) [8].

More recently, there has been an emergence of computational social science (CSS) which aims to analyze social phenomena through computational approaches [4]. Accompanying the abundance of digitized text data, one approach in CSS is to develop and utilize new algorithms and toolkits to conduct automated content analysis (e.g., [9]). Sentiment analysis, in particular, aims to automatically estimate or extract subjective sentiments that are expressed through texts, and it can be conducted at different levels. For instance, one may be interested in examining the overall sentiment from an article, or break down an article into sentences or words with their sentiments analyzed separately [7]. Methods that have been utilized to perform such analysis fall into two broad categories: dictionary-based and machine learning.

Dictionaries (or sometimes referred to as lexicons) contain a selected set of words with associated pre-defined sentiment scores. However, one drawback of this category of text analysis is that word order is often neglected. However, it may be crucial for certain types of sentiment analysis, especially at the aspect-level. For instance, an online restaurant review (e.g., from Yelp) may contain several aspects with different sentiments (e.g., good food but poor service) in a single sentence. Ignoring word order may not accurately estimate these aspect-based sentiments. Machine learning methods overcome this issue by encoding and processing texts in a sequential manner, where useful information about word order can be retained. As a result, the performance of machine learning models are often superior to dictionaries, as observed in several studies (e.g., [11]). Deep learning models in particular, have gained popularity in terms of ABSA over recent years [6]. Much work has been carried out to develop new models focusing on improving their predictive power, but they often fall short of advancing theory in explaining and understanding why and how people produce sentiments. It is our aim in this work to demonstrate how these predictive models can be used to investigate salient factors driving peoples' sentiments and link with existing results from qualitative studies, using online restaurant reviews as a case study.

Another approach in CSS is using agent-based modeling (ABM) to simulate complex systems by modeling individual agents and their interactions. In recent years, there have been a trend of integrating machine learning algorithms in and for agent-based models [5]. For example, dimensionality reduction and clustering algorithms have been utilized to analyse model outcomes. Machine learning models have also been used to train on human behavior data and subsequently represent agents during model executions. Agents may collaboratively optimize context-specific goals under the reinforcement learning framework. In order to answer our research question, we combine these two strands of CSS to explore how to utilize deep learning techniques to inform an agent-based model of consumer choices. In what follows, we briefly describe our methodology (Section 2) before presenting the results (Section 3). Finally, Section 4 proved a summary of the paper and identifies areas of further work.



(a) Total number of restaurant reviews by year. (b) Number of restaurant reviews in selected cities.

■ **Figure 1** Temporal (a) and spatial (b) distributions of restaurant reviews.

2 Methodology

The Yelp dataset is publicly available at <https://www.yelp.com/dataset> and contains more than 6 million text reviews on over 150,000 businesses in the United States from 2005 to 2021. In this study, we focus on a sampled set of reviews on restaurants prior to the outbreak of COVID-19. Our rationale for this is that COVID-19 has substantially altered consumer behaviours with respect to visiting restaurants. Figure 1a shows the total number of reviews over time while Figure 1b shows the number of reviews in representative cities.

We use a NLP approach to conduct ABSA on these Yelp restaurant reviews. Following the standard text pre-processing procedures in NLP, sampled texts are pre-processed to remove stop words, punctuation, and so on. Next, we draw on theories from computational linguistics and machine learning (e.g., [1]) to extract and categorize salient aspect terms from the text reviews, such as food, service, price, and ambiance, by applying a pre-trained language model from the PyABSA framework [13]. We also assign sentiments (i.e., positive, negative, and neutral) to these categorized aspects. Finally, a linear regression model is used to identify common sentiment patterns and examine how these patterns vary across different aspects.

While we can estimate casual effects of choice factors on star ratings through theoretical and statistical models, the results are only at an aggregate level, and it is difficult to gain insights on how they may differ for different consumer segments. This is mainly due to the lack of individual data which is a consequence of privacy and ethical concerns. As such, we turn to agent-based modeling to simulate artificial, heterogeneous agents (i.e., consumers) and their restaurant visiting patterns. During model initialization, restaurants are created with locations using information from Yelp, along with their average sentiment scores from the pre-trained language model mentioned above. Consumer agents are created at random places with a random attribute (i.e., student, middle-aged, or senior), which subsequently determines their restaurant preferences. For example, student agents are more sensitive to the price factor, whereas senior agents prefer restaurants with higher ambiance score, following findings from past studies through surveys and interviews [8]. At each step, consumer agents are informed by the NLP model results and visit the best restaurant based on their preferences. We also implement a null model in which consumer agents make random decisions on which restaurant to visit. An overview of the model logic is shown in Figure 2.

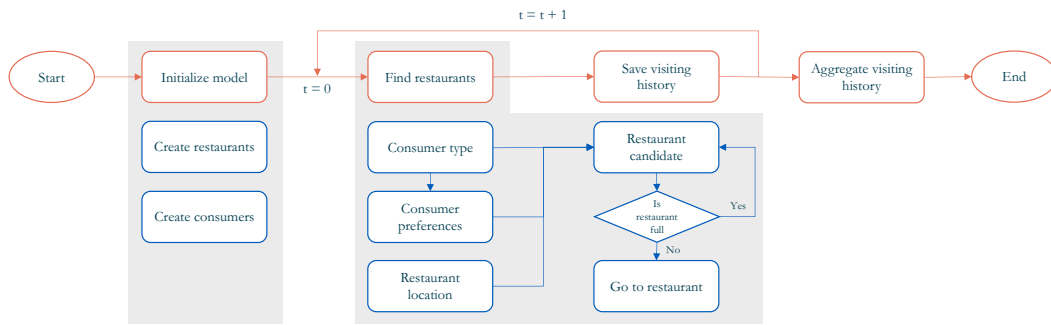


Figure 2 An overview of proposed agent-based model logic.

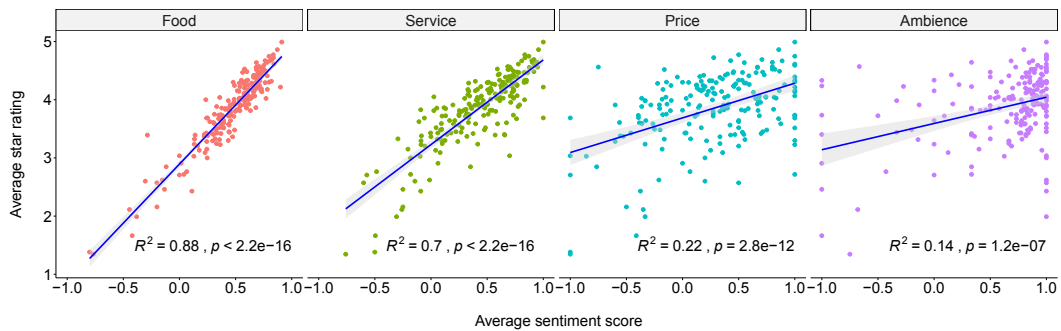


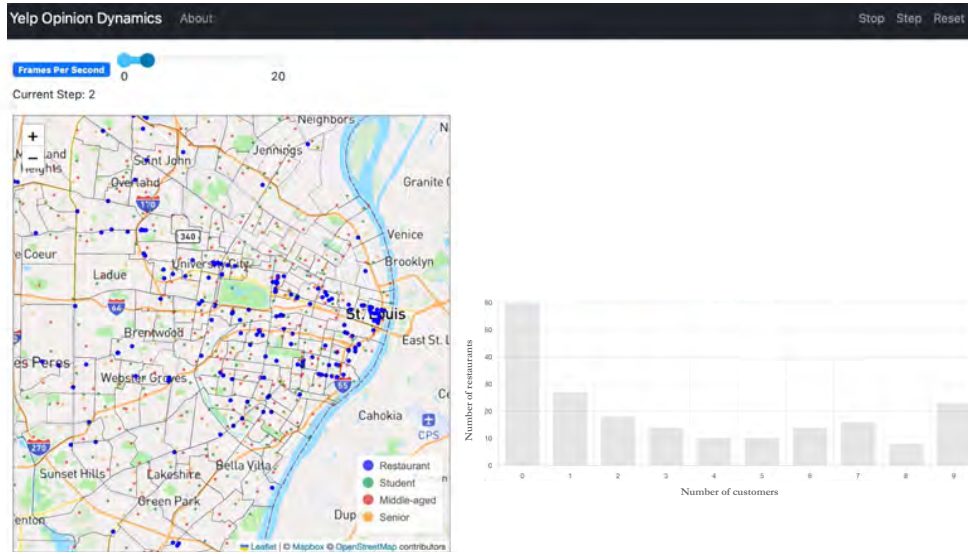
Figure 3 Average star rating vs. average sentiment by aspect category for 200 randomly selected restaurants in the City of St. Louis, MO.

3 Results

The pre-trained language model [13] transforms each text review and assigns a value to each main aspect category: food, service, price, and ambience, ranging from negative (-1), neutral (0), to positive (1). Although each text review from Yelp also includes a star rating ranging from 1 to 5, this information is never used during the model training. That is, model results are purely based on text data. To help answer RQ 1, we use star ratings as a proxy for consumer satisfaction, and plot average sentiment score of each aspect category against star ratings, for 200 randomly selected restaurants in the City of St. Louis, MO. The results are shown in Figure 3. Unsurprisingly, food appears to be the choice factor that is most strongly correlated with average star ratings while ignoring all other factors. This validates, and at the same time is validated by, previous qualitative studies through surveys and interviews [8]. Notably all results in Figure 3 are statistically significant ($p < 0.001$).

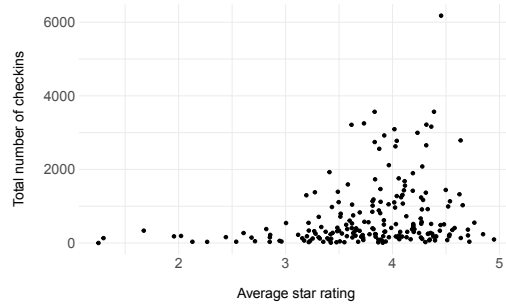
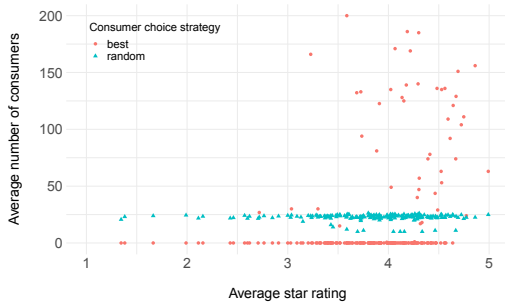
To understand consumers' decision-making processes and address RQ 2, we create a prototype model for the City of St. Louis using the Mesa framework in Python and its GIS extension Mesa-Geo [12]. A screenshot of this prototype model is shown in Figure 4a. Figure 4b shows the average results of 100 simulation runs, and compares it to the null model. Figure 4c shows the actual number of check-ins from Yelp. However a direct comparison between check-ins and visits is difficult to make because not all Yelp users do check-ins on each visit. Our model shows that there are significantly more consumer visits to restaurants with a star rating above 3 than to those with a star rating below 3. For the null model, restaurant visits are evenly distributed regardless of star ratings, which is to be expected. To some extent, this is also reflected in the actual number of check-ins versus actual star ratings.

Agent-Based Modeling



32

(a) Screenshot of the graphical user interface of the proposed agent-based model.



(b) Average number of consumers vs. star rating results from the simulations.

(c) Actual number of check-ins vs. star rating from the Yelp data.

■ **Figure 4** The prototype agent-based model (a) with simulated (b) and actual visiting patterns (c).

4 Summary and Areas of Further Work

Online customer reviews can provide valuable insights into various aspects of people’s dining experience, such as the quality of food and service. In this paper we have utilized ABSA methods on the Yelp dataset to extract and categorize reviewers’ opinions on restaurants in urban areas. These estimated opinions form a basis for subsequent statistical analysis and simulation through agent-based modeling. Within the context of this paper we see several areas of further work. A potential area to be further improved regarding sentiment analysis is to experiment with alternative language models of higher predictive performance, and fine-tune such models with more restaurant review data. In terms of the agent-based model, there is always room to extend and refine them. The first relates to incorporating more census data into the model when initializing the consumer agents in order to better stylize and build our synthetic population. It would also be interesting to explore a more finer time granularity that would capture different parts of the day such as mornings, afternoons and evenings as this might also impact visitation to different types of restaurants. Lastly, efforts could be made to calibrate model parameters and validate model results with restaurant

check-in data from Yelp or other check-in data sets such as Google. Even with these areas of further work, this paper demonstrates how using deep learning techniques can help advance our understanding of people's choices when it comes to visiting various locations within a city and how such analysis can be incorporated within agent-based models to explore how people interact with places and influence each other.



References

- 1 Eman Saeed Alamoudi and Norah Saleh Alghamdi. Sentiment classification and aspect-based sentiment analysis on yelp reviews using deep learning and word embeddings. *Journal of Decision Systems*, 30(2-3):259–281, 2021. doi:10.1080/12460125.2020.1864106.
- 2 Susan Auty. Consumer choice and segmentation in the restaurant industry. *Service Industries Journal*, 12(3):324–339, 1992. doi:10.1080/02642069200000042.
- 3 Milos Bujisic, Joe Hutchinson, and Haragopal Parsa. The effects of restaurant quality attributes on customer behavioral intentions. *International Journal of Contemporary Hospitality Management*, 26(8):1270–1291, 2014. doi:10.1108/IJCHM-04-2013-0162.
- 4 Claudio Cioffi-Revilla. *Introduction to computational social science*. Springer, New York, NY, 2017. doi:10.1007/978-3-319-50131-4.
- 5 Andrew Crooks, Alison Heppenstall, Nick Malleson, and Ed Manley. Agent-based modeling and the city: A gallery of applications. In Wenzhong Shi, Michael F. Goodchild, Michael Batty, Mei-Po Kwan, and Anshu Zhang, editors, *Urban Informatics*, pages 885–910. Springer, Singapore, 2021. doi:10.1007/978-981-15-8983-6_46.
- 6 Hai Ha Do, Penatiyana WC Prasad, Angelika Maag, and Abeer Alsadoon. Deep learning for aspect-based sentiment analysis: a comparative review. *Expert Systems with Applications*, 118:272–299, 2019. doi:10.1016/j.eswa.2018.10.003.
- 7 Bing Liu. *Sentiment analysis: mining sentiments, opinions, and emotions*. Cambridge University, Cambridge, UK., 2015. doi:10.1017/CB09781139084789.
- 8 Caroline Opolski Medeiros, Elisabete Salay, et al. A review of food service selection factors important to the consumer. *Food and Public Health*, 3(4):176–190, 2013. doi:10.5923/j.fph.20130304.02.
- 9 Ross Petchler and Sandra González-Bailon. Automated content analysis of online political communication. In Stephen Coleman and Deen Freelon, editors, *Handbook of digital politics*, pages 433–450. Edward Elgar Publishing, Cheltenham, UK, 2015. doi:10.4337/9781782548768.00037.
- 10 Kisang Ryu and Heesup Han. Influence of the quality of food, service, and physical environment on customer satisfaction and behavioral intention in quick-casual restaurants: Moderating role of perceived price. *Journal of Hospitality & Tourism Research*, 34(3):310–329, 2010. doi:10.1177/1096348009350624.
- 11 Wouter Van Atteveltdt, Mariken ACG Van der Velden, and Mark Boukes. The validity of sentiment analysis: Comparing manual annotation, crowd-coding, dictionary approaches, and machine learning algorithms. *Communication Methods and Measures*, 15(2):121–140, 2021. doi:10.1080/19312458.2020.1869198.
- 12 Boyu Wang, Vincent Hess, and Andrew Crooks. Mesa-Geo: A GIS extension for the Mesa agent-based modeling framework in python. In *Proceedings of the 5th ACM SIGSPATIAL International Workshop on GeoSpatial Simulation, (GeoSim '22)*, page 1–10. Association for Computing Machinery, 2022. doi:10.1145/3557989.3566157.
- 13 Heng Yang and Ke Li. PyABSA: Open framework for aspect-based sentiment analysis. *arXiv preprint arXiv:2208.01368*, 2022. doi:10.48550/arXiv.2208.01368.


Harnessing the Sunlight on Facades – an Approach for Determining Vertical Photovoltaic Potential

Franz Welscher¹  

Institute of Geodesy, Graz University of Technology, Austria

Ivan Majic  

Institute of Geodesy, Graz University of Technology, Austria

Franziska Hübl  

Institute of Geodesy, Graz University of Technology, Austria

Rizwan Bulbul  

Institute of Geodesy, Graz University of Technology, Austria

Johannes Scholz  

Institute of Geodesy, Graz University of Technology, Austria

Abstract

The paper deals with the calculation of the photovoltaic potential of vertical structures. Photovoltaic systems are a core technology for producing renewable energy. As roughly 50% of the population on planet Earth lives in urban environments, the production of renewable energy in urban contexts is of particular interest. As several papers have elaborated on the photovoltaic potential of roofs, this paper focuses on vertical structures. Hence, we present a methodology to extract facades suitable for photovoltaic installation, calculate their southness and percentage of shaded areas. The approach is successfully tested, based on a dataset located in the city of Graz, Styria (Austria). The results show the wall structures of each building, the respective shadow depth, and their score based on a multi-criteria analysis that represents the suitability for the installation of a photovoltaic system.

2012 ACM Subject Classification Applied computing → Mathematics and statistics; Applied computing → Environmental sciences

Keywords and phrases Vertical Photovoltaics, Facades, Southness, Multi-Criteria-Analysis, Shadow

Digital Object Identifier 10.4230/LIPIcs.GIScience.2023.82

Category Short Paper

Funding The presented results were obtained within the project PV4EAG (888491) funded by the Österreichische Forschungsförderungsgesellschaft (FFG). www.ffg.at.

1 Introduction

Renewable energy resources are crucial for sustainable development and mitigating the impact of climate change to achieve carbon neutrality. Solar energy is the largest inexhaustible source of clean energy in the world [12]. Solar photovoltaic (PV) systems play a vital role in harnessing solar energy and account for 10% of the world's electricity in the year 2022 [13].

The utilization of solar energy requires the assessment of the solar energy potential that quantifies the physically available solar radiation on the earth's surface [1]. Of special interest is the assessment of solar energy potential in urban environments, as almost 50% of the world's population currently lives in urban areas. Buildings are key structures in urban settings and understanding their solar potential is essential for the optimal utilization of solar

¹ corresponding author



energy in urban environments. Aside from rooftop [6] or ground mounted [8] PVs, there is an emerging interest in the solar potential of vertical structures (vertical PV) in recent years [4, 11]. Building facades provide additional space for the installation of PV systems that can complement the PVs installed on rooftops increasing the solar potential by 10-15% [2].

However, the assessment of the solar potential of building facades at a large scale (e.g. city level) is a challenging task. Most of the current solutions use complex radiation and shadow computation models that are computationally intensive and are not scalable from a few buildings to city level [2]. So, one of the major challenges for the assessment of vertical PV potential is the reasonable computational time [12]. Thus, computing the solar potential of building facades in urban areas needs to address the following challenges:

1. *Extraction of Facades*: Facades are vertical surfaces that show discontinuities in elevation models. Extracting vertical surfaces (facades) from elevation models is a complicated task. In addition, all facades of a building will not receive the same amount of solar radiation.
2. *Shadow Analysis*: Mutual shadowing due to the surrounding environment (other neighboring buildings, trees, etc) will reduce the solar radiation and this temporally varies.

In this paper, we address the aforementioned challenges by providing a novel approach for the assessment of vertical PV potential. The solution is computationally viable as it uses simple radiation and shadow computation models. The work takes the following simplifications and assumptions. (1) Only direct sunlight is considered for the radiation model. (2) We only consider facades with a minimum height of 3m and an area over 50m².

This paper has the following major contributions. (1) It presents a novel approach for the assessment of the PV potential of building facades that uses an efficient technique for shadow computation without using sky maps. (2) It provides an analysis of the important features for the vertical PV including southness, shading, height and width of building facades.

The remaining sections of the paper are organized as follows. In section 2 we provide an overview of existing approaches for assessing PV potential. Section 3 explains our proposed approach. In section 4 we demonstrate our approach on a small dataset. Finally, section 5 concludes this paper with a discussion of the contributions and an outlook of future work.

2 Related work

It is possible to estimate PV potential with statistical data, as [7] did on a country level (Austria) by describing the feasible potential of facades including physical/theoretical, technical, economic, and ecological/social limitations. However, approaches on a building level are more precise, but require a spatial data basis such as light detection and ranging (LiDAR) [4, 10], CityGML [14], aerial photogrammetry [9], or cadastral data where 2D to 3D objects can be extracted. E.g., a combination of LiDAR data, 2D, and 2.5D cadastral data was used by Desthieux et al. [4] to analyze the PV potential on rooftops and facades.

To handle the third dimension, shadow casting and solar radiation modeling were applied to 3D hyper-points covering the facade areas grid-like. Similar point-cloud-based methods were used in [11, 10], by calculating the sky view factor values for each point of a facade. The density of the points determines the accuracy and computational costs. Our approach uses 3D building data for vertical PV estimation too but it avoids computationally demanding point-cloud-based methods. This makes it applicable to large areas. We also introduce the southness indicator for building facades which was not addressed by literature yet.

3 Methodology

The approach proposed in this paper determines the PV potential of vertical surfaces (e.g. buildings facades). The developed workflow uses a raster containing shadow depth information to determine the PV potential of vertical structures (e.g. buildings). The shadow depth describes the height of a shadow cast by surrounding structures (e.g. trees, etc.).

The developed approach consists of six major steps. (1) Split into segments, (2) orientate geometries, (3) compute southness, (4) compute the wall area, (5) map shadow raster to wall segments and (6) the multi-criteria-analysis (MCA).

In more detail this means (1) we split buildings into wall segments and (2) orientate their geometries to ensure the outside of the building is to the left of the wall. (3) We then use this orientation to determine the southness of the wall segments, which is a value derived from the geographic direction that the outside of the wall segments is facing (see Section 3.2). (4) Based on the height and the length of the wall segments, their areas are derived. (5) To determine the shadowed area of the facade we map the shadow depth raster to vertical walls. This is done by computing the mean shadow depth in front of the wall segment (see Section 3.3). Consequently, the sun area of the facade is calculated by subtracting the shadowed area from the overall facade area. We also derive the percentage of the wall that is shadowed or non-shadowed from these values. Finally, (6) the PV potential of the walls is determined with an MCA. Scores and weights are assigned to the attributes *percentage of sun area*, *normalized southness indicator*, and *facade area* (see Section 3.4).

3.1 Preparation of vertical structures

Calculating vertical PV potential for a high quantity of buildings over large datasets can be computationally demanding and can be simplified by applying additional measures to the buildings. Firstly, instead of analyzing individual buildings, they can be dissolved into larger blocks. We propose a weighted average height for the dissolved building blocks. The weight is derived from the ratio of the area of the individual building to the total area of the block.

Secondly, once the buildings are dissolved into larger blocks they can be simplified further with line simplification algorithms such as the Douglas-Peucker algorithm [5]. This algorithm simplifies lines by excluding points based on a distance threshold.

3.2 Southness of the facades

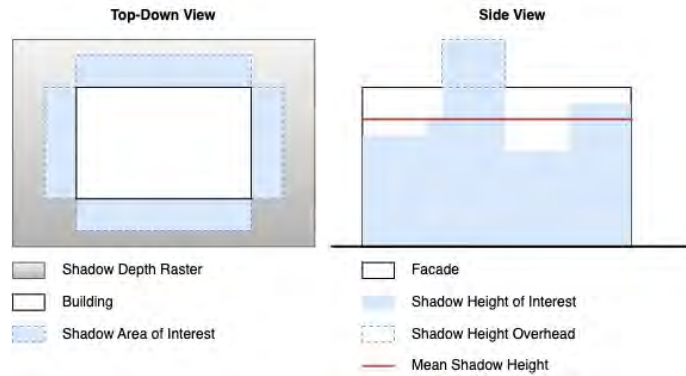
One of the most important features that determines how much sunlight exposure a facade will get is its orientation. The facade facing south is likely to receive more sunlight than the one facing north, while facades facing east and west are expected to fall in between. To capture and quantify this feature, we have come up with a normalized southness indicator that ranges from 0 for north-facing facades to 1 for south-facing facades and graduates equally between them regardless if the facade is facing east or west. Firstly, the azimuth of the facade is reduced to a value between 0° and 180°, and this value is then normalized. The reduction process differs depending on the azimuth and the exact formulae are shown in Table 1.

3.3 Mapping 2D-shadow raster to vertical walls

One step of assessing the potential of vertical areas is gaining information on the area covered by shadow. We map a 2D-Shadow-Raster to the wall segments by (1) buffering the outside of the walls, (2) intersecting the buffered wall with the shadow-raster and (3) deriving the mean shadow depth in front of the wall from the mean of the intersecting pixel values. The mean shadow depth is a negative value thus it is inversed to receive the mean shadow height.

■ **Table 1** Calculation of the normalized southness indicator depending on the azimuth of the line.

Azimuth [α]	$\alpha < 90^\circ$	$90^\circ \leq \alpha < 180^\circ$	$180^\circ \leq \alpha < 270^\circ$	$270^\circ \leq \alpha < 360^\circ$
Reduced Azimuth [α_r]	$90^\circ - \alpha$	$\alpha - 90^\circ$	$\alpha - 90^\circ$	$180^\circ - (\alpha - 270^\circ)$
Southness Indicator			$\alpha_r/180^\circ$	



■ **Figure 1** The top-down view shows the buffers in front of the walls that mark the shadow area of interest. The side view shows the different shadow heights in front of the facade covering up the facade area. To compute the mean shadow height, the shadow height must be reduced to the facade height to avoid shadow overhead.

We compute the shadow depth raster with an adjusted version of the QGIS Terrain shading plugin [3]. It uses the vertical angle and the horizontal angle (azimuth) of the sun to compute the shadow depth over the elevation model. We use an average shadow depth raster built from 12 rasters which represent 4 different days of the year with 3 times around noon for each day.

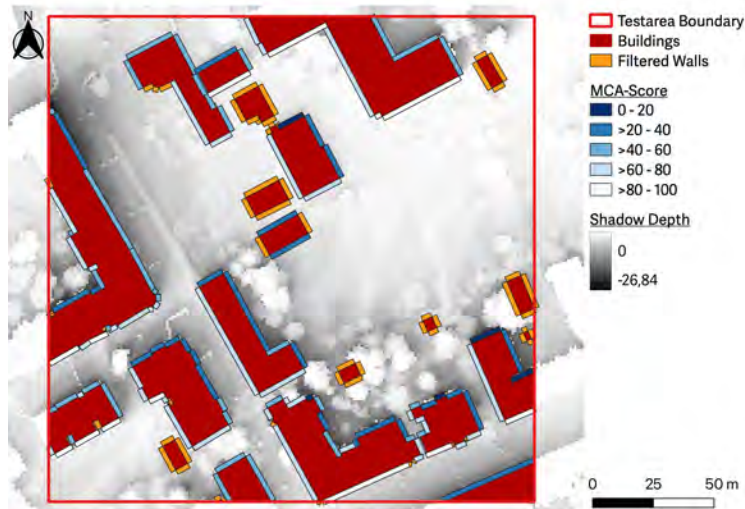
Figure 1 shows the problem at hand from two perspectives. To avoid a false mean shadow height we reduce the intersected height values to the wall’s height. Without this, the shadow height overhead can cause mean shadow height to exceed the wall height. By multiplying the mean shadow height with the length of the facade, the shadowed area is determined.

The wall buffer which is meant to capture the area in front of the building can intersect with the roof of the building itself or nearby treetops, leading to false shadow depth values. This is solved by filtering false shadow depth values with a threshold as they are considerably lower than the correct shadow depth values due to the height of these structures.

3.4 Multi-Criteria-Analysis

We derive the attributes that play a key role in the assessment of the PV potential of vertical structures by interviewing stakeholders and experts. Aside from binary filtering attributes like a *minimum height* or a *minimum area*, we determine the *percentage of sun area*, the *normalized southness indicator*, and the *total area* as properties for the assessment of the PV potential. With the *percentage of sun area* being the non-shadowed part of the facade area.

For the MCA it is required to assign the same score range to each property. Further, the three properties *percentage of sun area*, *normalized southness indicator* and *total facade area* are assigned weights based on their importance. This importance is gained from the knowledge of stakeholders and experts. The sum of all weights must be 1. By multiplying the weights with the assigned scores and building the sum of the weighted scores a total score is obtained that describes the PV potential of the vertical structure.



■ **Figure 2** The 200m by 200m test area consisting of 217 walls of which 117 walls are suitable for PVs depending on their MCA score.

■ **Table 2** The number of walls and the mean of the MCA attributes, as well as the mean total score by direction.

	Count	Mean Southness Indicator	Mean Sun Area Percentage (%)	Mean Total Score
North	35	0.16	43	39.76
West/East	40	0.43	47	48.19
South	42	0.79	74	73.82

4 Experiment

We test the proposed approach on a set of buildings in Graz, Styria (Austria). The test area is selected based on the even distribution of north, east/west and south facing facades (see 2). It has a size of $200m \times 200m$. It consists of 63 buildings, which we split into 217 walls. Of these walls, we determine 117 walls suitable for PVs based on the binary attributes. These are a minimum height of 3m and a minimum facade area of $50m^2$. By applying weights to the scores of the three properties we receive an overall PV suitability score ranging from 0 - 100. The used weights are 0.7 for *percentage of sun area*, 0.2 for *normalized southness indicator*, and 0.1 for *total facade area*.

Figure 2 visualizes the result for the test area. It shows the walls filtered by the binary attributes and the ones suited for PVs colored by their MCA score. It is visible that southward-facing walls tend to have a higher suitability score, than the ones facing northwards or east/west. A look at table 2 supports this. With an average total score of 73.82 southern walls tend to have the highest suitability, while northern walls have the lowest with 39.76. One aspect of interest is that the average percentage of the non-shadowed areas between north-facing walls and east-/west-facing walls only differs by 4%. As expected south-facing walls have the highest percentage of sun-covered area with a value of 74%.

5 Discussion and Outlook

In this paper, we present a novel approach for determining the PV potential of vertical surfaces. Further, we discuss the necessary pre-processing steps for the building data at hand. We highlight the normalized southness indicator as an attribute of assessing the PV potential of vertical surfaces. Additionally, we discuss our approach for determining the shadowed / non-shadowed area of vertical surfaces. The normalized southness indicator as well as the percentage of non-shadowed surface area play a vital role in multi-criteria-analysis of the PV potential of the vertical surfaces.

The key contributions of this paper are (1) the approach of assessing the PV potential of vertical surfaces, (2) the analysis of the relevant measures and indicators for vertical PV potential such as southness, shading, height and width of the facades. Further, (3) the findings may support the energy transition by finding potential facades for renewable energy production, and (4) it enables stakeholders and administration to make informed decisions concerning vertical PV areas which is of particular interest in urban contexts.

Future research aspects consist of evaluating the performance of the proposed approach by comparing it to existing methodologies. Further, it could be extended with additional attributes such as window surface area.

References

- 1 Athanasios Angelis-Dimakis, Markus Biberacher, Javier Dominguez, Giulia Fiorese, Sabine Gadocha, Edgard Gnansounou, Giorgio Guariso, Avraam Kartalidis, Luis Panichelli, Irene Pinedo, et al. Methods and tools to evaluate the availability of renewable energy sources. *Renewable and sustainable energy reviews*, 15(2):1182–1200, 2011.
- 2 Miguel Centeno Brito, Paula Redweik, Cristina Catita, Sara Freitas, and Miguel Santos. 3d solar potential in the urban environment: A case study in lisbon. *Energies*, 12(18):3457, 2019.
- 3 Zoran Cuckovic. Terrain shading: a qgis plugin for modelling natural illumination over digital terrain models, 2021. URL: <https://github.com/zoran-cuckovic/QGIS-terrain-shading>.
- 4 Gilles Desthieux, Claudio Carneiro, Reto Camponovo, Pierre Ineichen, Eugenio Morello, Anthony Boulmier, Nabil Abdennadher, Sébastien Dervev, and Christoph Ellert. Solar energy potential assessment on rooftops and facades in large built environments based on lidar data, image processing, and cloud computing. methodological background, application, and validation in geneva (solar cadaster). *Frontiers in Built Environment*, 4, 2018. doi:10.3389/fbui.2018.00014.
- 5 D Douglas and T Peuker. Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. *The International Journal for Geographic Information and Geovisualization*, 10:112–122, 1973.
- 6 Elham Fakhraian, Marc Alier, Francesc Valls Dalmau, Alireza Nameni, and José Casañ Guerrero. The urban rooftop photovoltaic potential determination. *Sustainability 2021, Vol. 13, Page 7447*, 13:7447, July 2021. doi:10.3390/SU13137447.
- 7 Hubert Fechner. Ermittlung des Flächenpotentials für den Photovoltaik-Ausbau in Österreich. *Studie im Auftrag von Österreichs Energie*, pages 1–69, 2020.
- 8 Christian Mikovits, Thomas Schauppenlehner, Patrick Scherhauser, Johannes Schmidt, Lilia Schmalzl, Veronika Dworzak, Nina Hampl, and Robert Gennaro Sposato. A spatially highly resolved ground mounted and rooftop potential analysis for photovoltaics in austria. *ISPRS International Journal of Geo-Information*, 10, 2021. doi:10.3390/IJGI10060418.
- 9 Arnadi Murtiyoso, Mirza Veriandi, Deni Suwardhi, Budhy Soeksmantono, and Agung B Harto. Automatic Workflow for Roof Extraction and Generation of 3D CityGML Models from Low-Cost UAV Image-Derived Point Clouds, 2020. doi:10.3390/ijgi9120743.

- 10 Jesús Polo, Nuria Martín-Chivelet, Miguel Alonso-Abella, and Carmen Alonso-García. Photovoltaic generation on vertical façades in urban context from open satellite-derived solar resource data. *Solar Energy*, 224(June):1396–1405, 2021. doi:10.1016/j.solener.2021.07.011.
- 11 P. Redweik, C. Catita, and M. Brito. Solar energy potential on roofs and facades in an urban landscape. *Solar Energy*, 97:332–341, 2013. doi:10.1016/j.solener.2013.08.036.
- 12 Naveed Rehman. *Solar Energy Potential Assessment On Façades Using Geo-referenced Digital Elevation Models*. PhD thesis, Auckland University of Technology, 2021.
- 13 REN21. Renewables 2022 global status report. Technical report, REN21 Secretariat, 2022.
- 14 Laura Romero Rodríguez, Eric Duminil, José Sánchez Ramos, and Ursula Eicker. Assessment of the photovoltaic potential at urban level based on 3D city models: A case study and new methodological approach. *Solar Energy*, 146:264–275, 2017. doi:10.1016/j.solener.2017.02.043.

Betweenness Centrality in Spatial Networks: A Spatially Normalised Approach

Christian Werner   

Department of Geoinformatics, University of Salzburg, Austria

Martin Loidl   

Department of Geoinformatics, University of Salzburg, Austria

Abstract

Centrality metrics are essential to network analysis. They reveal important morphological properties of networks, indicating e.g. node or edge importance. Applications are manifold, ranging from biology to transport planning. However, while being commonly applied in spatial contexts such as urban analytics, the implications of the spatial configuration of network elements on these metrics are widely neglected. As a consequence, a systematic bias is introduced into spatial network analyses. When applied to real-world problems, unintended side effects and wrong conclusions might be the result. In this paper, we assess the impact of node density on betweenness centrality. Furthermore, we propose a method for computing spatially normalised betweenness centrality. We apply it to a theoretical case as well as real-world transport networks. Results show that spatial normalisation mitigates the prevalent bias of node density.

2012 ACM Subject Classification Theory of computation → Theory and algorithms for application domains; Mathematics of computing → Graph algorithms; Mathematics of computing → Network flows; Information systems → Geographic information systems

Keywords and phrases spatial network analysis, edge betweenness centrality, flow estimation, SIBC, spatial interaction, spatial centrality, urban analytics

Digital Object Identifier 10.4230/LIPIcs.GIScience.2023.83

Category Short Paper

Supplementary Material *Other (Code and examples)*: <https://doi.org/10.5281/zenodo.8125632>

1 Introduction

Betweenness centrality is a key metric for assessing node and edge importance in networks. It is based on computing the share of shortest paths that pass each edge or node in relation to the total number of paths in a network. Thereby it reveals the relative importance of edges or nodes for enabling interaction within the network. While centrality metrics can generally be applied to spatial networks, many complex effects occur that are still to be fully understood [2]. As direct consequence of the definition of betweenness centrality, the number of nodes, their location, and their morphological embedment within the network determine centrality. The key aspect to be questioned within spatial applications is the assumption that each pair of nodes has equal influence on centrality. This characteristic implies that the spatial density of nodes strongly influences betweenness centrality.

State of the Art

Due to the generic network science origin of centrality metrics, their focus lies on topological rather than spatial properties of networks. However, important steps for integrating spatial aspects into centrality concepts have been accomplished e.g. by considering the spatial length of edges and paths in betweenness centrality. Further research assessed how different forms of spatial networks influence centrality and how such networks can be characterised through



© Christian Werner and Martin Loidl;
licensed under Creative Commons License CC-BY 4.0

12th International Conference on Geographic Information Science (GIScience 2023).

Editors: Roger Beecham, Jed A. Long, Dianna Smith, Qunshan Zhao, and Sarah Wise; Article No. 83; pp. 83:1–83:6

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

the use of centrality metrics [2]. Various studies applied betweenness centrality for assessing spatial transport networks - for Indian railways[6], Paris and London transport[4] or urban road networks in Germany[5]. All these examples have in common that they do not consider the impact of spatial configuration on centrality metrics, thus neglecting its potential bias.

Application-driven research proposed domain-specific concepts for weighting origin-destination pairs in computation of centrality, which may mitigate bias of spatial configuration. In urban analytics and transport modelling, weighting based on flow estimation is common. Spatial interaction between origins and destinations is modelled, resulting in an estimate for travel demand. Despite the long history of such methods, adequate modelling is highly complex and has been found not to meet real-world observed patterns in many cases.

Research Gap

While numerous studies applied betweenness centrality to spatial networks, effects of the spatial configuration of nodes on centrality have not been regarded systematically. Furthermore, we identified the lack of a simplistic null model of betweenness centrality for applications in spatial networks that avoids introducing complex (behavioural) models.

To fill this gap, we first assess the problem in more detail and then provide a method to compensate the influence of spatial node density. Motivated from the application domain of urban analytics and mobility, we focus on edge betweenness centrality as known key measure from which node betweenness centrality may easily be derived [2]. Where helpful, we motivate our theoretical considerations using examples of real-world transport networks.

2 Method

The Problem Illustrated

To illustrate the impact of node density on edge betweenness centrality (c_B), we use a simplistic reference case. In a network constructed as a regular grid, c_B is known to be highest in the spatial centre, as shown in figure 1 a). If we subdivide one grid cell by adding an additional node per edge and one node at the cell centre, we observe a shift in high c_B towards the newly subdivided area, visualised in figure 1 b). One may think of this as a city block to which access paths for pedestrians have been added. While the overall structure of this virtual residential area remained the same (i.e. no buildings have been added or removed), centrality shifted significantly. This can be explained by the fundamental definition of c_B . As we introduce new nodes - in the given case five nodes are added to a cell originally consisting of four nodes - each of these new nodes introduces an equally important origin and destination for all shortest paths computation. As a consequence, the influence of paths from and to this cell increases in relation to all paths within the given network.

To quantify this gain in influence, we can calculate the change in contribution of paths from and to the given cell relative to all paths within the network. Following the definition of c_B (see equation 1), it is more precisely the number of origin-destination relations that start or end within the given cell that we are interested in. As known from normalisation of c_B , the total number of origin-destination relations in a directed graph consisting of n nodes is $n(n-1)$. As one single node has the role as origin as well as destination for o-d relations to all other nodes, it contributes $2(n-1)$ o-d relations. Consequently, we can express the relative contribution of one node to all possible relations as $\frac{2(n-1)}{n(n-1)}$. In a more generic form, we can quantify the contributed relations of i nodes to the network beyond the given cell as $\frac{2i(n-1)-i(i-1)}{n(n-1)}$.

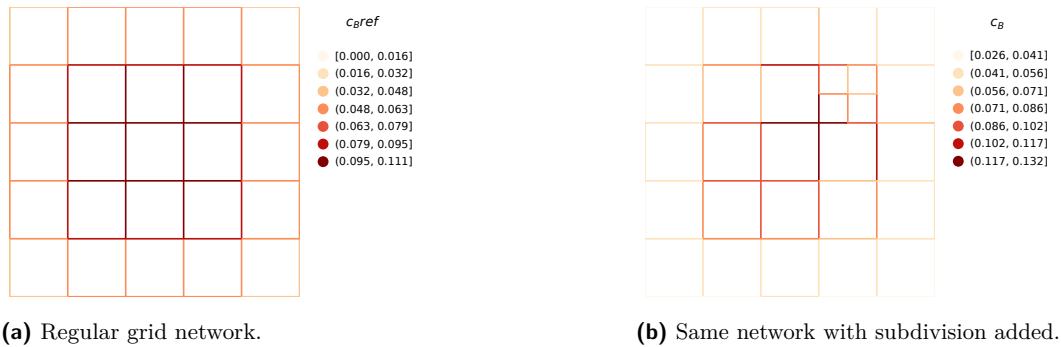


Figure 1 The problem illustrated: Influence of node density on centrality.

Considering the case of adding subdivision to a regular grid, we can assess the contribution I of nodes forming a given cell C and the subdivided cell C' as follows:

$$I_C = \frac{2a(n-1) - a(a-1)}{n(n-1)} \text{ and } I_{C'} = \frac{2(m+a)(n+m-1) - (m+a)(m+a-1)}{(n+m)(n+m-1)}$$

where a denotes the number of nodes originally constituting the cell and m denotes the number of nodes added through subdivision.

The relative increase in influence of cell C can be assessed as $\frac{I_{C'}}{I_C}$ which for $a \ll n$ and $m \ll n$ can be approximated as: $n^2(a+m) \frac{-2n}{-2an^3} = \frac{a+m}{a}$. Thus, the influence of the given cell on centrality increases approximately proportional to the increase in node count for the same area for commonly large networks.

For our minimalistic example with $n = 36; a = 4; m = 5$, the cell's influence increases by $\frac{81}{205} / \frac{67}{315} \approx 1.858$. Due to the small network size and relatively high a and m , this value does not reach the approximate value for large networks of $\frac{4+5}{4} = 2.25$.

To summarise this section, we were able to show that the influence of spatial node density on betweenness centrality can be assumed to be proportional for commonly large networks. Consequently, spatial variation in node density has significant impact on betweenness centrality. While this might be intended in specific application cases, it appears unintended for generic, unbiased assessments and remains hard to control for in general.

Proposed Method: Spatial Betweenness Centrality

To mitigate the effect of varying node density on betweenness centrality, we propose a method for computing spatially normalised betweenness centrality. We refer to edge betweenness centrality as c_B and to our proposed spatial edge betweenness centrality as c_{SB} . The main idea behind c_{SB} is to weight all paths contributed to centrality per origin-destination pair relative to the area covered by their origin and destination nodes.

Our proposed method consists of two steps: 1) determining the spatial coverage per node, and 2) computing spatially weighted centrality based on node coverage.

Determining the spatial coverage per node. Spatial coverage of nodes can be determined using tessellation of the network space. While tessellation using Voronoi polygons is common e.g. for retrieving a network null model, it does not render suitable in the given case. Especially in networks with high variability in edge length, Voronoi polygons may intersect non-adjacent edges. Furthermore, motivated from the mobility domain, we assume that interaction is generated along edges rather than at nodes. Therefore, we propose utilising an edge-based tessellation such as the method described by Araldi and Fusco using proximity

bands [1] or network-based Thiessen tessellation. The area of each polygon then describes the spatial coverage per edge. Each node's spatial coverage can thus be derived as the total spatial coverage of all edges adjacent to a node divided by two.

Compute spatially weighted centrality based on node coverage. Edge betweenness centrality $c_B(e)$ is defined for a graph $G(V, E)$, where V refers to the set of nodes and E refers to the set of edges as follows:

$$c_B(e) = \sum_{s,t \in V; s \neq t} \frac{\sigma_{s,t}(e)}{\sigma_{s,t}} \quad (1)$$

In this formula, for any pair of origin s and destination t nodes, $\sigma_{s,t}$ denotes the number of shortest paths from s to t , and $\sigma_{s,t}(e)$ refers to the quantity of paths that pass edge e .

In standard c_B , each origin-destination pair contributes equally to centrality. Consequently, the weight each o-d relation contributes equals to one: $w_{relation}(s, t) = 1$. For spatial normalisation we want to weight the paths contributed to centrality per o-d pair proportionally by their origin and destination node spatial influence (weights). Therefore, we propose a weight function that distributes an origin node's spatial influence (area covered) to all other nodes proportionally to their relative spatial influence as a destination:

$$w_{relation}(s, t) = w(s) \frac{w(t)}{\sum_{u \in V; u \neq s} w(u)} \text{ for } s, t \in V; s \neq t$$

where $w(v)$ refers to the weight of a node v , respectively its spatial coverage. The full definition of our proposed spatial betweenness centrality metric consequently reads as:

$$c_{SB}(e) = \sum_{s,t \in V; s \neq t} \frac{\sigma_{s,t}(e)}{\sigma_{s,t}} w(s) \frac{w(t)}{\sum_{u \in V; u \neq s} w(u)} \quad (2)$$

In order to obtain normalised values for c_{SB} , the absolute values are divided by the total area covered by origin nodes: $c_{SBnorm}(e) = c_{SB}(e) / \sum_{v \in V} w(v)$.

We propose an implementation based on spatial interaction incorporated betweenness centrality (SIBC)[7]. It builds upon Brandes algorithm [3] and adds a weight function $f(s, t)$, which represents a measure of spatial interaction - known flow or estimated flow based on a gravity model [7]. If one pre-computes the o-d weight matrix, it can be employed as spatial interaction matrix in the SIBC method. For applicability in large networks, we suggest computing o-d weights stepwise per origin, alongside solving the single-source shortest path (SSSP) problem.

3 Results

In this section we provide centrality assessments for different networks using both, standard edge betweenness centrality c_B and our proposed spatial variant c_{SB} .

The artificial case: regular grid network. As first example we assess the network that we used to illustrate the problem in section 2.

In figure 1 we can observe that c_B shows a shift of high centrality towards the subdivided cell, whereas such a shift is not present in c_{SB} shown in figure 2 a) and b). The differences between c_{SB} for the subdivision case and c_B for the regular grid case are relatively small. In contrast, figure 2 c) highlights the mitigated shift of high centrality when applying c_{SB} .

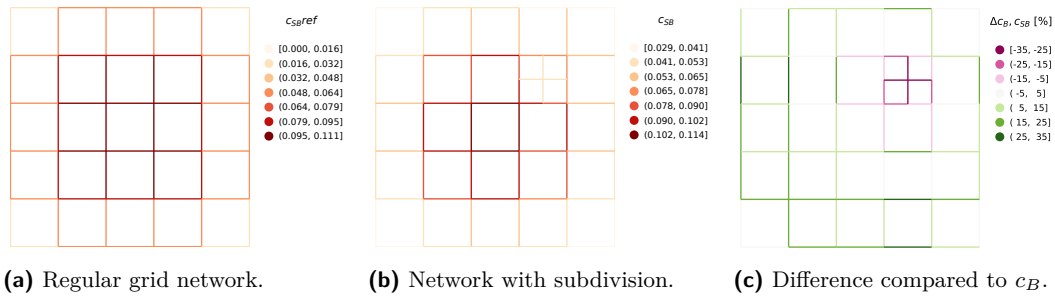


Figure 2 Spatial betweenness centrality applied to the original problem case.

The real-world case: street networks. Additionally, we computed c_B and c_{SB} for several extracts of real-world road networks of varying form. For brevity, we only present one example here and provide more cases online at <https://doi.org/10.5281/zenodo.8125632>.

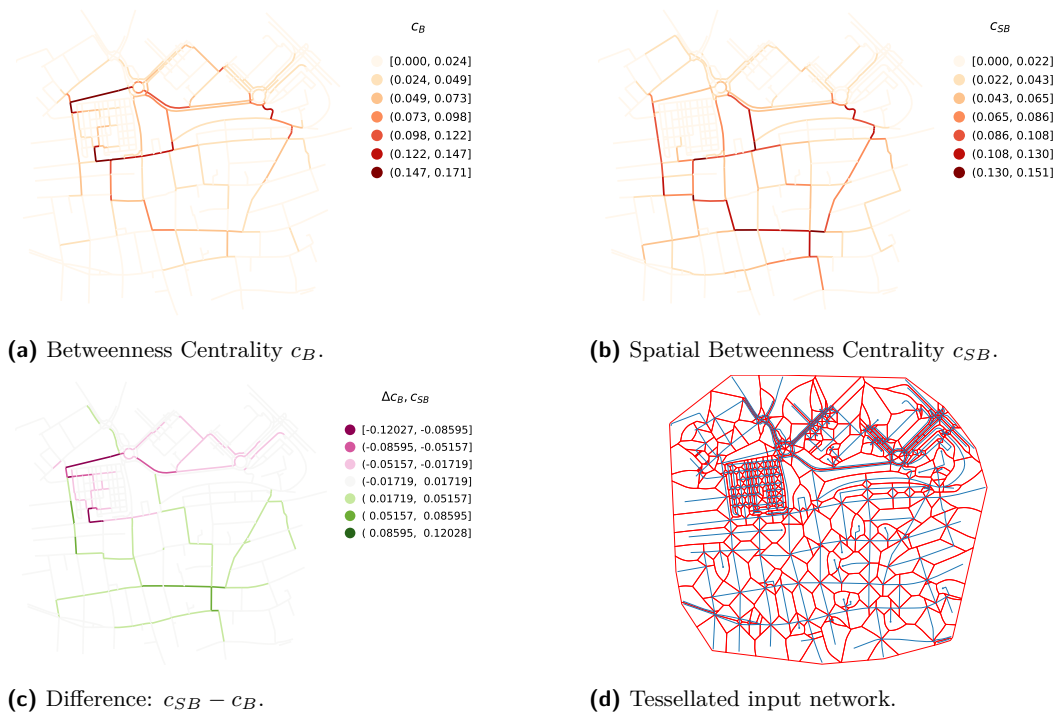


Figure 3 Betweenness centrality for a subset of a real-world street network (Stuttgart, Germany).

Results for the real-world case using a subset of Stuttgart, Germany are presented in figure 3. High node density is present in the North, whereas lower node density is prevalent in the centre and South, which is visible in subfigure d). Accordingly, the size of tessellation polygons decreases with higher node density. When comparing subfigures a) and b) or assessing the differences in subfigure c), c_B puts a clear emphasis on routes linking the high-density areas. For c_{SB} , part of these links also show above-average centrality. However, additional links in the centre and South are more pronounced in c_{SB} .

For all networks assessed, we can observe a tendency of higher centrality values for c_B in proximity of areas with higher node density compared to spatially normalised centrality c_{SB} .

4 Discussion and Outlook

We showed that spatial variation in node density has significant influence on betweenness centrality for spatial networks. Unless node density is an intended indicator to consider in a specific application case, we regard this as systematic bias that needs to be addressed. E.g. for applications in mobility, standard betweenness centrality c_B may only render suitable results, if node density spatially correlates with population density.

With the concept of spatial betweenness centrality c_{SB} we propose a generic solution that utilises spatial normalisation to weight the contribution of individual relations. Results applying c_{SB} show a clear mitigation of bias introduced through variations in node density in c_B . Spatial betweenness centrality c_{SB} can therefore provide a generic null model of betweenness centrality in spatial networks.

For practical application of c_{SB} , edge effects need to be considered. One may utilise a network covering larger extent than the area of interest for assessment to allow shortest paths on edges outside the area of interest and to avoid edge effects in tessellation. Future research should shed more light on specific edge effects of c_B and c_{SB} .

Depending on the domain-specific application case, additional factors may be integrated into c_{SB} assessments. Non-uniform weight may be applied to areas of e.g. different land use. This also allows for excluding certain areas from contributing to centrality computation as origin and destination. Furthermore, combination e.g. with population data can open new application scenarios.

We see great potential in the use of spatial betweenness centrality c_{SB} for unbiased, generic assessment of spatial networks. It combines both, morphological properties with spatial embedment of the network. However, it may depend on the specific domain application, whether c_{SB} or an advanced domain-specific modelling approach is preferable.

References

- 1 Alessandro Araldi and Giovanni Fusco. From the street to the metropolitan region: Pedestrian perspective in urban fabric analysis. *Environment and Planning B: Urban Analytics and City Science*, 46(7):1243–1263, September 2019. doi:10.1177/2399808319832612.
- 2 Marc Barthelemy. *Morphogenesis of Spatial Networks*. Lecture Notes in Morphogenesis. Springer International Publishing, Cham, 2018. doi:10.1007/978-3-319-20565-6.
- 3 Ulrik Brandes. A faster algorithm for betweenness centrality. *The Journal of Mathematical Sociology*, 25(2):163–177, June 2001. doi:10.1080/0022250X.2001.9990249.
- 4 L. da F. Costa, B. A. N. Travençolo, M. P. Viana, and E. Strano. On the efficiency of transportation systems in large cities. *Europhysics Letters*, 91(1):18003, July 2010. doi:10.1209/0295-5075/91/18003.
- 5 Stefan Lämmer, Björn Gehlsen, and Dirk Helbing. Scaling laws in the spatial structure of urban road networks. *Physica A: Statistical Mechanics and its Applications*, 363(1):89–95, April 2006. doi:10.1016/j.physa.2006.01.051.
- 6 Parongama Sen, Subinay Dasgupta, Arnab Chatterjee, P. A. Sreeram, G. Mukherjee, and S. S. Manna. Small-world properties of the Indian railway network. *Physical Review E*, 67(3):036106, March 2003. doi:10.1103/PhysRevE.67.036106.
- 7 Xiaohuan Wu, Wenpu Cao, Jianying Wang, Yi Zhang, Weijun Yang, and Yu Liu. A spatial interaction incorporated betweenness centrality measure. *PLOS ONE*, 17(5):e0268203, May 2022. doi:10.1371/journal.pone.0268203.

Predicting visit frequencies to new places

Nina Wiedemann¹  

Institute of Cartography and Geoinformation, ETH Zürich, Switzerland

Ye Hong  

Institute of Cartography and Geoinformation, ETH Zürich, Switzerland

Martin Raubal  

Institute of Cartography and Geoinformation, ETH Zürich, Switzerland

Abstract

Human mobility exhibits power-law distributed visitation patterns; i.e., a few locations are visited frequently and many locations only once. Current research focuses on the important locations of users or on recommending new places based on collective behaviour, neglecting the existence of scarcely visited locations. However, assessing whether a user will return to a location in the future is highly relevant for personalized location-based services. Therefore, we propose a new problem formulation aimed at predicting the future visit frequency to a new location, focusing on the previous mobility behaviour of a single user. Our preliminary results demonstrate that visit frequency prediction is a difficult task, but sophisticated learning models can detect insightful patterns in the historic mobility indicative of future visit frequency. We believe these models can uncover valuable insights into the spatial factors that drive individual mobility behaviour.

2012 ACM Subject Classification Information systems → Geographic information systems; Computing methodologies → Neural networks; Applied computing → Transportation; Information systems → Location based services

Keywords and phrases Human mobility, Visitation patterns, Place recommendation, Next location prediction

Digital Object Identifier 10.4230/LIPIcs.GIScience.2023.84

Category Short Paper

Supplementary Material *Software (Source code)*: <https://github.com/mie-lab/predict-visits> archived at `swh:1:dir:c0c080878ee26ac806daac00fd25458dbfeb5406`

Text (Implementation details): https://github.com/mie-lab/predict-visits/blob/main/supplementary_information.pdf, archived at `swh:1:cnt:5eb98f0df940d22a9e3b0a1dbf883bef1b029688`

1 Introduction

Large-scale tracking data collected from mobile phone users are crucial for location-based services such as place recommendations [8]. One field of research is the so-called next location² prediction, which is concerned with finding the immediate next location an individual will visit [7]. Such predictions could be used for recommendations, navigation advice or on-demand transport services. The developments in this field, however, suffer from the heavy-tailed distribution of visit frequencies [3]; i.e., many locations are visited only once and are thus difficult if not impossible to predict [11]. Specifically, Cuttone et al. [1] find that 70% of locations are visited only once, and 20-25% of the visits are to new locations. The interest of users in these locations is primarily assessed upfront via recommendation systems that

¹ Corresponding author

² Since the term "place" describes the subjectively experienced form of a geographic location [19], we will use the term "location" throughout the paper to objectively denote a user's activity cluster.



leverage insights from aggregated user behaviour, e.g., the general popularity of a place. Many systems were developed for this purpose, mainly based on data from location-based social networks (LBSN), and employ (context-aware) collaborative filtering [9, 2, 18]. For a recommender system to successfully suggest entirely new locations to the user, data from many users in the same region must be available, which is often hampered due to the sensitive nature of tracking data [4].

At the same time, the mobility of a *single* user already allows one to draw insights about a user’s interest in new locations. For example, the spatial layout [10] and topology of the mobility behaviour [20, 16], or the category frequencies of the user’s previously visited locations [17] can help to estimate the spatial distribution of future visitation patterns [11]. In light of this possibility, we argue for a new problem formulation: *Predicting the frequency of future visits to a newly visited location, given the historic mobility of a single user*. In other words, assuming that we observe a user visiting a location for the first time, can we predict whether they will return to this location, in a scenario where knowledge about collective mobility patterns is scarce? We argue that this problem is mistaken as a subtask of recommender systems or next location prediction. In contrast, it requires special modelling approaches to learn efficiently from individual historic mobility patterns. Successful approaches could decide whether a location will become part of a user’s activity set [6] and possibly unveil hidden patterns in the user’s location preferences. Moreover, the gained knowledge will support the online next location prediction that needs to consider new locations at runtime, or improve individualized transport recommendation and planning.

In this paper, we formalize our new problem termed “visit frequency prediction”, and present an approach to frame it as a supervised learning task. We experiment with self-attention-based and graph-based neural network models to efficiently process the historical tracking data. As expected, predicting the visit frequency to new locations is challenging due to the lack of information about the user’s motives for visiting the location. Nevertheless, we find that neural network models can find patterns in the historic mobility that are predictive of future visits, improving over the baseline methods.

2 Problem formulation

Let $L^u = \{l_1^u, \dots, l_m^u\}$ denote the set of all locations visited by user u . A location is defined by point coordinates or an area, where the user performed some stationary *activity* (e.g., working or catering). Locations can be derived, e.g., by clustering GNSS data or from check-ins to known POIs in LBSN. In practice, a user visits these locations sequentially, represented as a list S^u of n visit events, for example, $S^u = [l_2^u, l_1^u, l_2^u, l_4^u]$ with $n = 4$. The visit frequency $\nu(l)$ is thereby the number of visits to location l , e.g., $\nu(l_2^u) = 2$ in the example. Let $S_{i:j}$ be the excerpt of the chain from the i -th until the j -th element in S (excluding the j -th). We assume that at a specific point t , we observe that a new location $l_\theta^u \notin S_{1:t}^u$. The task is to predict $\nu(l_\theta^u)$ in $S_{t:n}^u$ given the historic mobility $S_{1:t}^u$.

One potential approach is to train a model to learn a mapping g such that, optimally, $\nu(l_\theta^u) = g(S_{1:t}^u, l_\theta^u, u)$, where the model could leverage 1) feature representation of the previously visited locations $f(l)$, $l \in S_{1:t}^u$ and the visit frequency $\nu(l)$ of these locations, 2) user characteristics u , and 3) features of the new location $f(l_\theta^u)$. Note that the model can be fitted to the data of many users, but, at inference time, it should be possible to apply the model to the data from a single user, potentially in a different geographical region.

3 Methods

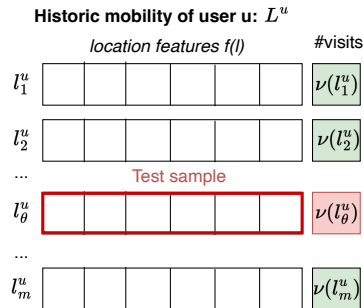


Figure 1 Approaching the visit frequency prediction problem as a supervised task.

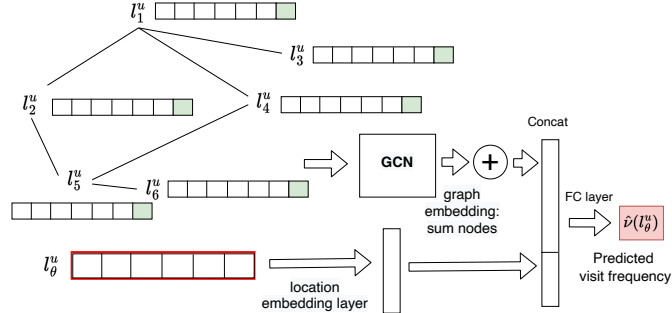


Figure 2 Graph-based model for learning visit frequency of new locations from the historic location graph. The graph is embedded and concatenated with the new location's features.

We propose a supervised approach to tackle the visit frequency prediction problem. In each training step, one location l_θ^u is removed from the user's overall tracking data (see Figure 1). The pruned mobility data $L^u \setminus l_\theta^u$ and $S^u \setminus l_\theta^u$ (i.e., the historic mobility, pretending that l_θ^u was never visited), as well as features of the removed location $f(l_\theta^u)$ are provided as input, and the visit frequency $\nu(l_\theta^u)$ is the desired output. We utilize the following features as f : The projected coordinates of l relative to the home location, the location purpose encoded as a one-hot vector, the average start hour of visits to l , and POI features. This leads to a vector of 24 entries. We implement a simple median and a k-nearest neighbor (kNN) approach as baselines and then test a fully connected neural network (MLP), a multi-head self-attention (MHSA) model, and a graph convolutional network (GCN) on the task. Each model is described in the following. For implementation details, see our code and supplementary material available at <https://github.com/mie-lab/predict-visits>.

The simple median baseline is given by $\hat{\nu}(l_\theta^u) = \text{median}(\{\nu(l^u) \mid l^u \in L^u\})$. This approach yields the same output for all queried locations of a user. For a more informed baseline, we consider a kNN approach, estimating the unknown visit frequency as $\hat{\nu}(l_\theta^u) = \frac{1}{k} \sum_{l \in N(l_\theta^u)} \nu(l)$, where $N(l_\theta^u)$ is the set of k nearest neighbors of l_θ^u in $L^u \setminus l_\theta^u$. We measure the distance between locations by the Euclidean distance of their feature vectors f .

For the MLP and the MHSA model, we provide a fixed set of m locations from the historic mobility of a single user, $L^u \setminus l_\theta^u$, and the new location l_θ^u as input. We hypothesize that the locations with the highest activity are most predictive of the visit frequency to new locations, and therefore select the m locations with the highest visit frequency. They are sorted by the frequency and are featurized by f , leading to an input matrix of size $(m + 1) \times 24$. The matrix is flattened to be fed into the model. The MLP is a simple fully-connected two-layer network, whereas our MHSA follows the implementation by Hong et al. [5] for location prediction. A graph approach, on the other hand, allows for a variable number of input locations per user. Our approach is shown in Figure 2. The graph is passed through a Graph-Resnet [13], and the node embeddings are combined with average pooling, yielding a single vector of fixed size. This graph embedding is then concatenated with the embedding of the new location features $f(l_\theta^u)$ passed through a single layer. The last layer yields the estimated visit frequency.

4 Results

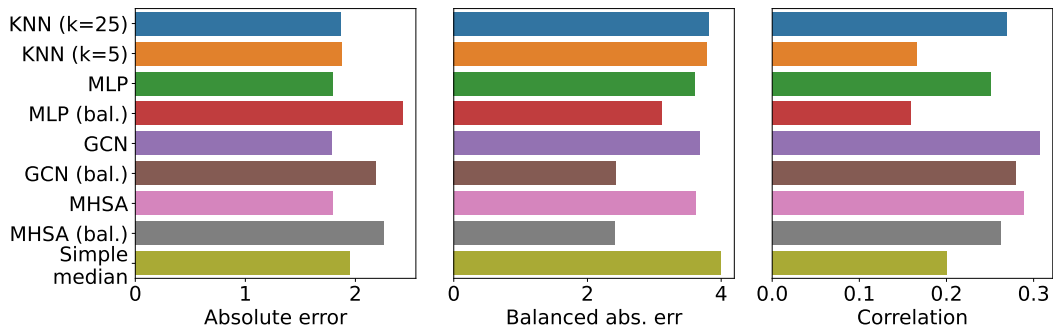
4.1 Data

We utilize high-quality and activity-labelled GNSS data from three tracking studies: Green Class 1 (GC1), Green Class 2 (GC2) [12] and yumuv [15]. All three studies were executed in collaboration with the Swiss Federal Railways (SBB) and aimed to evaluate the impact of Mobility-as-a-Service offers. The participants were tracked via a GNSS-based app and were asked to manually label their activities. The app already preprocesses the raw GNSS track points by inferring stationary *staypoints* and continuous movement *triplegs*, which are further processed with the Python library Trackintel [14]. Trackintel derives a set of visited locations from a user’s tracking data using the DBSCAN clustering algorithm. After preprocessing, we included 139 users for GC1, 48 for GC2 and 653 for yumuv, who visited 104.5k, 35.7k and 127.3k distinct locations respectively. To align the tracking period, we split the data into time bins of three months. Finally, following Martin et al. [16], we transform the tracking data into a location graph for the GCN-based approach with the same hyperparameter setting. By the visit frequency prediction definition given above, the model should be applicable to unseen users in other geographic regions. Therefore, we split the data into train and test set on a *dataset*-level for the experiment; i.e., the train set \mathcal{D}_{train} comprises randomly sampled data from the GC1 and yumuv studies, and \mathcal{D}_{test} is sampled from GC2. To focus on rarely-visited locations, we only use locations that were visited up to ten times as test locations (l_θ). This cutoff on average excludes three locations per person.

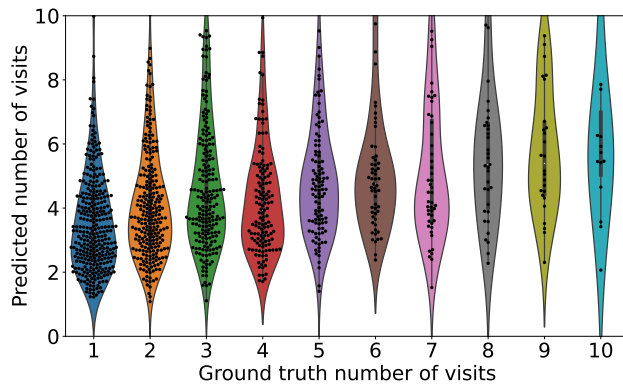
4.2 Model comparison

Figure 3 shows the results for all tested models. We first consider the mean absolute error (MAE), i.e. $\frac{1}{|\mathcal{D}_{test}|} \sum_{l_\theta^u \in \mathcal{D}_{test}} |\hat{\nu}(l_\theta^u) - \nu(l_\theta^u)|$. The absolute error is generally low (around 1.8) for all models, and complex models only improve marginally over the baselines. However, the MAE is misleading due to the imbalance between the visit frequencies: Many locations are visited only once, whereas very few are visited ten times. For a more insightful evaluation, we propose to consider the balanced MAE: $\frac{1}{10} \sum_{i=1}^{10} \left(\frac{1}{|\mathcal{D}_{test}^i|} \sum_{l_\theta^u \in \mathcal{D}_{test}^i} |\hat{\nu}(l_\theta^u) - \nu(l_\theta^u)| \right)$

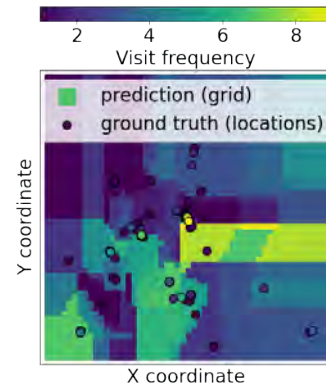
As Figure 3 (middle) shows, the balanced MAE is 3.99 for the simple median baseline and improves to 3.78 for the best kNN model. The neural network models yield a substantial improvement if they are also trained with *balanced* data (denoted by "bal." in Figure 3), meaning that the batches at train time were sampled such that each visit frequency from 1 and 10 appear equally often. The balanced GCN and balanced MHSA model yield the best performance with a balanced MAE of 2.43, indicating that these models can indeed learn patterns in historic mobility. The results for the balanced GCN are also visualized as a



■ **Figure 3** Model comparison on the visit frequency prediction problem for new users.



■ **Figure 4** Violinplot of visit frequency predicted by the GCN compared to the ground truth.



■ **Figure 5** Spatial distribution of predicted visit frequencies.

violin plot in Figure 4. While the test set is imbalanced and the predictions are very noisy, there is a clear shift in the distribution of predicted frequency with increasing ground truth visit frequency.

Finally, the Pearson correlation coefficient ρ of predicted and ground truth visit frequencies of the test data is shown in Figure 3 (right). The GCN and MHSA models again achieve the best performance with ρ up to 0.3. In general, the results indicate that predicting visit frequencies to newly visited locations is a difficult task. The value of the predicted frequencies for real applications is limited so far, even though they are more accurate than the baselines.

5 Discussion and outlook

The increasing availability of user location data gives rise to new research opportunities in the context of location recommendation and prediction. We have introduced a new problem that, for the first time, regards the importance of newly visited locations by approximating their projected visit frequency. Our preliminary results show that the task suffers from similar difficulties as next location prediction, namely noisy data, lack of information and inherent stochasticity in user decisions. The difficulty is also due to the strong imbalance of the ground-truth visit frequency. However, other models or additional context data may improve performance.

A well-trained visit frequency prediction model could also be applied to map the probability of visits to new locations. This analysis would yield insights into the spatial distribution of visit frequencies learnt by the model. An example is shown in Figure 5, where we systematically sampled locations within the convex hull of the visited locations of one user. The heatmap of predicted visits is based on hidden patterns detected in the ground-truth visit frequencies (dots, locations that are only visited once are filtered out for visibility). An analysis of the spatial visitation patterns, e.g., with respect to the spatial layout and distances of frequently visited locations, may improve the understanding of user behaviour. Thus, we believe that visit frequency prediction is an exciting endeavour, and we hope that our problem formulation and preliminary methodology inspire further research on this topic.

References

- 1 Andrea Cuttone, Sune Lehmann, and Marta C González. Understanding predictability and exploration in human mobility. *EPJ Data Science*, 7:1–17, 2018.

- 2 Tuan Hung Dao, Seung Ryul Jeong, and Hyunchul Ahn. A novel recommendation model of location-based advertising: Context-Aware Collaborative Filtering using GA approach. *Expert Systems with Applications*, 39(3):3731–3739, 2012.
- 3 Marta C Gonzalez, Cesar A Hidalgo, and Albert-Laszlo Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008.
- 4 Linrui Han. Personal Privacy Data Protection in Location Recommendation System. In *Journal of Physics: Conference Series*, volume 2138, page 012026, 2021.
- 5 Ye Hong, Henry Martin, and Martin Raubal. How do you go where?: improving next location prediction by learning travel mode information using transformers. In *Proceedings of the 30th International Conference on Advances in Geographic Information Systems (SIGSPATIAL '22)*, 2022.
- 6 Ye Hong, Henry Martin, Yanan Xin, Dominik Bucher, Daniel J. Reck, Kay W. Axhausen, and Martin Raubal. Conserved quantities in human mobility: From locations to trips. *Transportation Research Part C: Emerging Technologies*, 146:103979, 2023.
- 7 Ye Hong, Yatao Zhang, Konrad Schindler, and Martin Raubal. Context-aware multi-head self-attentional neural network model for next location prediction. *arXiv preprint arXiv:2212.01953*, 2022.
- 8 Haosheng Huang, Georg Gartner, Jukka M Krisp, Martin Raubal, and Nico Van de Weghe. Location based services: ongoing evolution and research agenda. *Journal of Location Based Services*, 12(2):63–93, 2018.
- 9 Defu Lian, Yong Ge, Fuzheng Zhang, Nicholas Jing Yuan, Xing Xie, Tao Zhou, and Yong Rui. Content-Aware Collaborative Filtering for Location Recommendation Based on Human Mobility Data. In *2015 IEEE International Conference on Data Mining*, pages 261–270, 2015.
- 10 Bin Liu, Yanjie Fu, Zijun Yao, and Hui Xiong. Learning geographical preferences for point-of-interest recommendation. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '13)*, pages 1043–1051, 2013.
- 11 Xin Lu, Erik Wetter, Nita Bharti, Andrew J Tatem, and Linus Bengtsson. Approaching the Limit of Predictability in Human Mobility. *Scientific reports*, 3(1):2923, 2013.
- 12 Henry Martin, Henrik Becker, Dominik Bucher, David Jonietz, Martin Raubal, and Kay W. Axhausen. Begleitstudie SBB Green Class - Abschlussbericht. *Working Paper No. 1439, Institute for Transport Planning and Systems, ETH Zürich*, 2019.
- 13 Henry Martin, Dominik Bucher, Ye Hong, René Buffat, Christian Rupprecht, and Martin Raubal. Graph-ResNets for short-term traffic forecasts in almost unknown cities. In *NeurIPS 2019 Competition and Demonstration Track*, pages 153–163, 2020.
- 14 Henry Martin, Ye Hong, Nina Wiedemann, Dominik Bucher, and Martin Raubal. Trackintel: An open-source Python library for human mobility analysis. *Computers, Environment and Urban Systems*, 101:101938, 2023.
- 15 Henry Martin, Daniel Reck, Kay Axhausen, and Martin Raubal. Empirical use and impact analysis of MaaS. Technical report, ETH Zurich, 2021.
- 16 Henry Martin, Nina Wiedemann, Daniel J Reck, and Martin Raubal. Graph-based mobility profiling. *Computers, Environment and Urban Systems*, 100:101910, 2023.
- 17 Seyyed Mohammadreza Rahimi and Xin Wang. Location Recommendation Based on Periodicity of Human Activities and Location Categories. In *Advances in Knowledge Discovery and Data Mining: Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD '13)*, pages 377–389, 2013.
- 18 Sulis Setiowati, Teguh Bharata Adji, and Igi Ardiyanto. Context-based awareness in location recommendation system to enhance recommendation quality: A review. In *2018 International Conference on Information and Communications Technology*, pages 90–95, 2018.
- 19 Yi-Fu Tuan. *Space and place: humanistic perspective*. Springer, 1979.
- 20 Nina Wiedemann, Henry Martin, and Martin Raubal. Unlocking social network analysis methods for studying human mobility. *AGILE: GIScience Series*, 3:19, 2022.

Waffle Homes: Utilizing Aerial Imagery of Unfinished Buildings to Determine Average Room Size

Carson Woody   

Human Geography Group, Oak Ridge National Laboratory, TN, USA

Tyler Frazier 

Human Geography Group, Oak Ridge National Laboratory, TN, USA

Abstract

A primary function of the Population Density Tables Project (PDT) at Oak Ridge National Laboratory is to produce residential population densities per 1000 sq. ft. for each country and their associated first-level administrative units. This is accomplished by utilizing the average size of different types of dwelling areas (urban, rural, single-family, multi-family, etc.) and the average household size provided by a country's Census or statistical bureau records. This data is available for the majority of Europe, North America, and large swathes of Asia, but is less easily found in Africa and South America. In these regions, Censuses generally report dwelling area by number of rooms, which poses the challenging question of how we can translate number of rooms to dwelling size when no dwelling size areas are available with which to compare. Using sub-meter resolution satellite imagery of Accra, Ghana, this challenge can be tackled using imagery of roofless buildings currently under construction that show the interior floor plan of the dwelling. A sample of buildings from the different neighborhoods of Accra can be digitized to provide an estimate and range of average room sizes of dwellings. This average room size can then be translated to a total dwelling area using the "number of rooms occupied by a household" variable from the Ghanaian Census. This intermediate step between average dwelling size and number of rooms occupied, fills the missing link that prevents PDT from continually producing new population densities for countries where dwelling size is unavailable through any official means.

2012 ACM Subject Classification Social and professional topics → Cultural characteristics

Keywords and phrases Urban Analytics, Aerial Imagery, Satellite Imagery, Population Density, Human Geography, Africa, Residential Dwellings

Digital Object Identifier 10.4230/LIPICs.GIScience.2023.85

Category Short Paper

Acknowledgements This manuscript has been authored by UT-Battelle, LLC under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes. The Department of Energy will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (<http://energy.gov/downloads/doe-public-access-plan>).

1 Introduction

Oak Ridge National Laboratory's Population Density Tables (PDT), is an information system with a graphical user interface that measures population density for over 60 facility types by people per 1000 sq. ft. [9]. Generally, this is performed under a Bayesian approach using



© Carson Woody and Tyler Frazier;
licensed under Creative Commons License CC-BY 4.0

12th International Conference on Geographic Information Science (GIScience 2023).

Editors: Roger Beecham, Jed A. Long, Dianna Smith, Qunshan Zhao, and Sarah Wise; Article No. 85; pp. 85:1–85:6

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany



■ **Figure 1** Waffle Home Example (Accra, Ghana) [4].

prior knowledge from expert elicitations and from gathering new data on individual facility types in a region to update the model [10]. Uncertainty is propagated throughout each model, starting at the data entry stage, where a range of data can be input into the PDT interface. Using this approach, PDT estimates ambient building occupancy at the national and subnational level. To model residential building occupancy, PDT uses the RevengC R package to reverse engineer census data and produce an uncensored table of population density probabilities. This R package is used in PDT in the form of the Census Tool, which uses average household size and average dwelling size as inputs to produce a residential dwelling density per 1000 sq. ft. These inputs are primarily obtained from country statistical bureau produced censuses or statistical documentation as PDT views Census data as the most authoritative sampling source [2].

Dwelling sizes are generally accounted for in national censuses and state statistical bureaus in the form of floor area in sqm. or sq. ft. However, in the majority of Africa and South America, recent censuses account for dwelling size in the form of “number of rooms occupied” in a dwelling unit. Without knowing the actual size of these rooms occupied, it poses a challenge to measure residential population density for a large portion of the world.

To use the data provided from censuses where dwellings are measured by number of rooms occupied as opposed to floor area, a link between the rooms inhabited and the actual size of these rooms is needed. We found this intermediary in Accra, Ghana (AOI), and its large volume of unfinished “honeycomb” structures where one can easily see the interior layout of the building as seen in Figure 1. These were coined “waffle homes” due to their grid iron like appearance, and along with Accra, were found in several other large African cities, like Lagos, Khartoum, and Conakry. From this type of imagery, one can easily identify the separate rooms in a building. Subsequently, individual rooms can be digitized using GIS software. With a statistically robust sample of digitizations, an average room size can be defined.

2 Related Works

Traditionally, the compound house is one of the most common typologies of housing in Accra, and tends to be the subject of research in this area. This is a common home style for lower income households and generally consist of a series of single hall units surrounding an open courtyard with shared kitchen and bathroom facilities, and generally cover an area of 100 sqm. [1]

However, compound homes have been losing popularity and are considered outdated as households prefer single-family living structures, which encourages builders to prioritize more single-family homes and apartments. This trend coupled with the continued influx of people to Ghana's urban areas, has increased the need for housing and rooms built, with an estimated 5.7 million rooms having been required to house the population by 2020 [6]. Many compound homes are not expected to survive the new wave of construction as new and affordable housing targets the middle class [3].

This wave of new building can be seen in the aerial imagery of Accra. Little research could be found using the aerial imagery of under construction homes to determine dwelling and average room sizing, but there has been research using architectural floor plans to study the general area of different types of rooms in different housing layouts [5]. This work relied on published floor plans from New Zealand to measure room area and went further by differentiating rooms by function and floor layout. Our study only relies on the aerial imagery of under construction dwellings without the ability to ascertain each room's individual function, but the New Zealand study did show that knowing the average area of different rooms of a home can lead to an accurate estimate of the home's total floor area [5]. While building typology differs greatly between New Zealand and Accra, the same idea of measuring room size can be used to find dwelling areas in our AOI.

3 Methodology

The Greater Accra Region, a subnational unit of Ghana including the country's capital, was used as the AOI for this new methodology. This region works as an ideal case study due to its lack of a census designated average dwelling floor area and its large number of waffle home type structures from which to sample. In addition, the 2021 Ghana Population and Housing Census recently became available which provided the necessary data on dwelling size by number of occupied rooms as well as average household sizes. This number of rooms occupied data provided by the 2021 Ghana Census will be used in conjunction with the waffle home digitizations to determine an average floor area [7].

Samples of waffle homes were gathered using Digital Globe/Maxar imagery from the Greater Accra Region. Samples were manually identified using imagery tiles from each district of the Greater Accra Region, and a total of 1267 sample points were identified by their unique grid-iron appearance in aerial imagery. A point data set was created from these structures, and the embedded PDT smart sampling tool was used to create a statistically robust selection of buildings across Accra. The embedded smart sampler tool works by randomly selecting points from the data set to digitize until the sample is large enough to closely represent the "true average" of the initial data set. It does this by having the data set pass three statistical tests before being considered a statistically robust selection of the data set. This was done due to time constraints and to lessen the number of buildings and rooms that needed to be digitized to create a data set of room sizes.

The first test the embedded smart sampler tool uses is a consistent distribution check, which ensures the data has the same estimated distribution, in this case a log-normal distribution. Then, the data set must pass a mean percentage change check, to ensure that

85:4 Waffle Homes

as each new data point is sampled, the average is moving closer to the true average. An acceptable mean percentage change was set at 5 percent for each time a new data entry is added, and once the entries fall within this threshold a specified number of times, it passes this test. The last test, a Margin of Error check, sets a maximum margin of error, here it is 10 percent, and the embedded tool will calculate a range of values using a confidence interval of 95 percent. Once each new entry falls within the margin of error range a specified number of times, it will pass test three.

A data set of 697 rooms was created using the smart sampler tool by randomly selecting from the 1267 waffle house points to digitize the areas of each defined room in the dwelling. The 2021 Ghana Census’s definition of an “occupied room” was used, which states that occupied rooms include living rooms, bedrooms, sleeping rooms, and dinning rooms, but excludes closets and bathrooms. This definition was used along with the 2018 Ghana Building Codes stipulation that an occupied room shall not be less than 9.5 sqm. to uniformly remove any bathrooms/closets from the room data set.

4 Results

This new data set of 697 rooms averages out to be 20.38 sqm. per room. Throughout this process, an additional 45 “waffle house” points were sampled and each of the rooms were digitized during further testing and research. This brought the total rooms data set up to 1089 rooms, and a new room average of 20.37 sqm. per room. Even with the addition of 393 rooms to the data set, the change in area was only .01 sqm. This supports the assumption that expanding the survey to additional points across the region would still return a similar average room size.

The resulting sample data set of waffle-home rooms shows a log-normal distribution with the majority of samples falling under 30 sqm., reaching a cumulative frequency of 85.49 percent at 30 sqm. The majority of the samples themselves fell between 10 to 20 sqm with 578 of the 1089 sample areas. Figure 2 shows the area distribution of the data set.

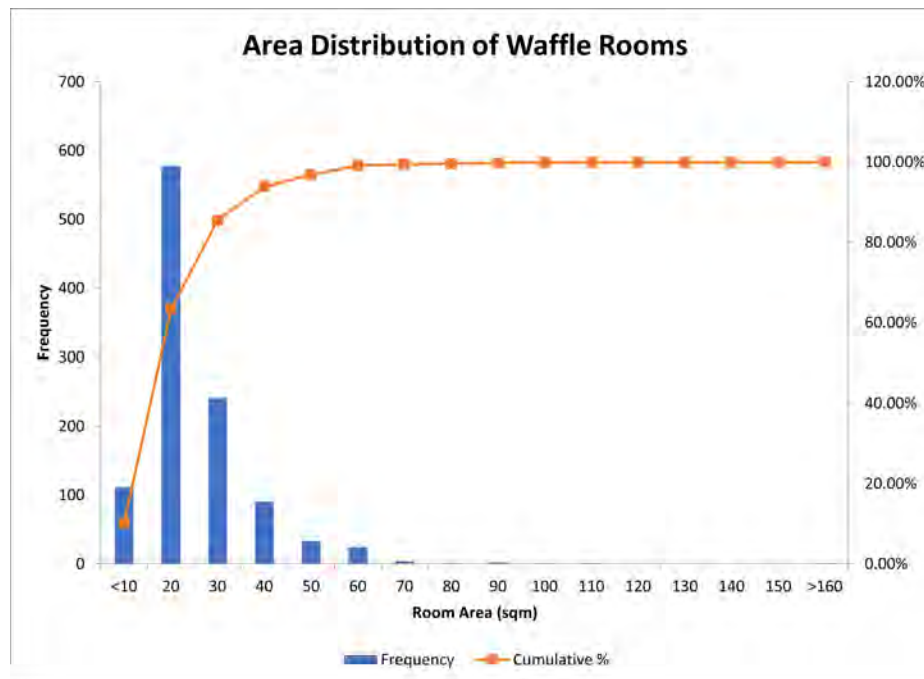
The Ghana Living Standards Survey of 2008 (GLSS5) was used as a comparison metric. While dated, it provides the average dwelling floor area for the Greater Accra area that the 2021 Ghana Population and Housing Census lacks [8]. The Ghana Population and Housing survey was used to determine the average area occupied by using the average number of rooms occupied (2.09) of the Greater Accra Region (GAMA), along with the newly determined average room size [7]. Table 1 shows a comparison of average areas between the two studies.

■ **Table 1** GLSS5 and Waffle Home Data set Comparison.

	GLSS5 Urban [8]	GLSS5 GAMA [8]	Waffle Homes GAMA
Average Room Size	19.59 sqm	23.66 sqm	20.37 sqm
Average Area Occupied	33.3 sqm	42.6 sqm	42.57 sqm

The results produced by the Waffle Home methodology are similar to those produced by the GLSS5. Validation is difficult in this scenario as there are no recent state produced dwelling area averages against which to compare for Ghana or from the neighboring countries. More data will be needed, either from a state sponsored survey or additional waffle home POIs, to provide a more rigorous validation.

PDT’s future goal is to build upon our current Bayesian modeling process to utilize this new method of sampling room floor areas along with the “rooms occupied” tables provided by state-level censuses to create total dwelling areas. PDT observation models capture



■ **Figure 2** Histogram and Cumulative Frequency of Waffle Home Data set.

uncertainty at the data entry level and propagate it through the model so it can be observed in the resulting probability density function. This feature will be necessary in the model expansion for this format of data to capture the uncertainty associated with using average room sizes and will likely utilize a range of average room sizes to better represent the dwelling areas of a region.

5 Discussion

This method's main limitation for future research is the time intensive process used to manually pinpoint and digitize the hundreds of individual rooms used to obtain the average room size. Ideally, this process would be able to expand to the entire country to create an Overall Residential population density number for all of Ghana, including the rural regions. Additionally, there is human error and bias associated with manually obtaining data points for digitization. We can address this by building onto one of our current PDT projects, which is a geospatial object detection tool, that will be trained to identify each instance of a waffle-home structure from the tiles of Digital Globe/Maxar imagery. This will create a data set of nearly every waffle home point in a country that can then be sampled in a similar manner as described in the methodology to find the average room size. This will ensure a wider spatial range of rooms identified to provide an average that better represents the wide spatial array of dwelling sizes. This image detection system can be easily implemented into existing imagery processing pipelines.

A secondary challenge involves the actual waffle home data points themselves. As seen from the literature, building priority has been placed on single-family and apartment style homes for a growing middle-class and less on compound style homes. As this study only samples buildings under construction from the past three years, there is a likelihood for the average room size to skew on the higher end. Samples were taken from more informal-type

settlements and rural areas, but the majority were drawn from new urban construction. The use of the object detection system would increase the overall sample size of the waffle homes. Using this expanded sample, we could ensure a better balance in samples from rural and urban settlements towards a more representative sample of waffle homes. This, along with a future emphasis to breakout residential densities into socio-economic classes, as PDT already incorporates in its modeling, will be used to ensure lower-income and informal housing is better represented in these estimates.

This new methodology of utilizing the aerial imagery of rapidly expanding construction and GIS software to measure room areas has the potential to fill in data gaps for large patches of the world where normal data collection methods are unavailable. There is a need to model residential population density across the world, not only to be able to model as much of the world as possible, but also to better apply aid in humanitarian crises and environmental disasters. Unique data collection methods are necessary in areas where available data is lacking the necessary information. While this method has a fairly simple but unconventional approach, it is important to find new methods of data acquisition to circumnavigate challenges in data gaps.

References

- 1 Lewis Abedi Asante, Emmanuel Kofi Gavu, Jonathan Zinzi Ayitey, and Alexander Sasu. The changing face of compound houses in ghana and its effect on rental value: A case study of selected neighborhoods in kumasi, ghana. Technical report, African Real Estate Society (AfRES), 2015.
- 2 Samantha Duchscherer, Robert Stewart, and Marie Urban. revengc: An R package to Reverse Engineer Summarized Data. *The R Journal*, 10(2):114–123, 2018. doi:10.32614/RJ-2018-044.
- 3 Yinka Ibukun. How city life transformed ghana’s compound houses. *Bloomberg: CityLab*, 2021. URL: <https://www.bloomberg.com/news/features/2021-05-06/the-design-history-of-ghana-s-compound-houses>.
- 4 Google Maps Imagery. Accra google maps imagery, 2023. Imagery ©2023 Google, Imagery ©2023 CNES / Airbus, Maxar Technologies, Map data © 2023. URL: <https://www.google.com/maps/place/Accra,+Ghana/@5.6250365,-0.1023598,198m/data=!3m1!1e3!4m6!3m5!1s0xfdf9084b2b7a773:0xbed14ed8650e2dd3!8m2!3d5.6037168!4d-0.1869644!16zL20vMGZueWM>.
- 5 Iman Khajehzadeh and Brenda Vale. Estimating the floor area of a house knowing its number of rooms and how these are named. In *Back to the Future: The Next 50 Years, (51st International Conference of the Architectural Science Association (ANZAScA))*, December 2017.
- 6 Dahlia Nduom. Housing and culture in ghana: A model for research and evidence-based design. *ARCC Conference Repository*, September 2018. URL: <https://www.arcc-repository.org/index.php/repository/article/view/533>.
- 7 Ghana 2021 Population and Housing Census. Ghana 2021 population and housing census: Housing characteristics. Technical report, Ghana Statistical Service, 2021.
- 8 Ghana Statistical Service. Ghana living standards survey, report of the fifth round (glss 5). Technical report, Ghana Statistical Service, 2008.
- 9 Robert Stewart, Marie Urban, Samantha Duchscherer, Jason Kaufman, April Morton, Gautam Thakur, Jesse Piburn, and Jessica Moehl. A bayesian machine learning model for estimating building occupancy from open source data. *Natural Hazards*, 2016. doi: 10.1007/s11069-016-2164-9.
- 10 Marie Urban, Robert Stewart, Scott Basford, Zachary Palmer, and Jason Kaufman. Estimating building occupancy: a machine learning system for day, night, and episodic events. *Natural Hazards*, 2023. URL: <https://link.springer.com/article/10.1007/s11069-022-05772-3#citeas>.

A Comparison of Global and Local Statistical and Machine Learning Techniques in Estimating Flash Flood Susceptibility

Jing Yao ✉ 

Urban Big Data Centre, School of Social and Political Sciences, University of Glasgow, UK

Ziqi Li ¹ ✉ 

Department of Geography, Florida State University, Tallahassee, FL, USA

Xiaoxiang Zhang ✉

Department of Geographic Information Science, College of Hydrology and Water Resources, Hohai University, Nanjing, China

Changjun Liu ✉

Department of Flood and Drought Disaster Reduction, China Institute of Water Resources and Hydropower Research, Beijing, China

Liliang Ren ✉

State Key Laboratory of Hydrology-Water Resources and Hydraulic Engineering, College of Hydrology and Water Resources, Hohai University, Nanjing, China

Abstract

Flash floods, as a type of devastating natural disasters, can cause significant damage to infrastructure, agriculture, and people's livelihoods. Mapping flash flood susceptibility has long been an effective measure to help with the development of flash flood risk reduction and management strategies. Recent studies have shown that machine learning (ML) techniques perform better than traditional statistical and process-based models in estimating flash flood susceptibility. However, a major limitation of standard ML models is that they ignore the local geographic context where flash floods occur. To address this limitation, we developed a local Geographically Weighted Random Forest (GWRF) model and compared its performance against other global and local statistical and ML alternatives using an empirical flash floods model of Jiangxi Province, China.

2012 ACM Subject Classification Computing methodologies → Machine learning

Keywords and phrases Machine Learning, Spatial Statistics, Flash floods, Susceptibility

Digital Object Identifier 10.4230/LIPIcs.GIScience.2023.86

Category Short Paper

Funding This work was supported by the National Key Research and Development Program of China [grant numbers No. 2019YFC1510601]; the Economic and Social Research Council, UK [grant number ES/P011020/1, ES/S007105/1].

1 Introduction

Flash floods are one of the most devastating natural disasters, which often occurs within a short period of time and can be caused by a variety of factors such as intense rainfall, rapid snow melt, landslides, and dam failure. Given their rapid speed and strong force, flash floods can cause significant damages to properties, infrastructures, and even loss of life. As a result, flash flood risk mitigation and management are of fundamental importance if sustainable

¹ Corresponding author



development is to be achieved. Flash flood susceptibility estimation has long been an effective means adopted by practitioners and policymakers to assist with development of flood risk reduction strategies, land use planning and emergency resource deployment [6] [8].

Common approaches that have been widely adopted in the estimation of flash flood susceptibility include statistical, hydrodynamic models and geographical information system (GIS) based spatial analyses. The examples of statistical models include regression analysis, frequency ratio, weights-of-evidence, and analytical hierarchy process, among others [6]. Hydrodynamic models usually predict the propensity of an area to flash floods by simulating the water flow during a rainfall event [13]. GIS-based approaches often combine potential factors that contribute to flash floods (e.g., rainfall topography and land use) to identify areas at risk, mainly utilizing remote sensing images [12].

In recent years, with the emergency of big data (e.g., weather and water levels) collected by various sensors as well as the advances in high-performance computing techniques, artificial intelligence (AI) particularly machine learning (ML) has been increasingly applied in evaluating and predicting flash flood susceptibility [9]. Common ML approaches such as support vector machine (SVM), random forest (RF), neural network (NN) have demonstrated better performance than traditional methods like statistical and hydrodynamic models [8] [2] [1]. However, a major limitation of existing ML approaches is that they ignore the geographic nature of flash floods. Often, the same set of hyperparameters are employed for all observations without considering the geographic context of each flash flood event. It is worth mentioning that there have been several recent developments in GeoAI that incorporate spatiality into modeling.

[3] developed Geographical Random Forests (GRF), in which a separate RF model is fitted for each location. One limitation of GRF is that, although it considers the local nature of the phenomenon, it does not allow geographical weighting in the training, which ignores the distance-decay effect for most geographical processes. [4] improved GRF to incorporate geographical weighting, but the prediction process for unseen data is less explicit and does not allow the weighting kernel to vary spatially. [5] developed a Geographically Weighted Neural Network (GWNN) model, in which geographical weighting is imposed on the loss function during model training. However, GWNN does not allow hyperparameters to vary spatially, thus failing to account for local variations in the underlying processes.

To this end, in this paper, we address limitations in recent GeoAI developments by allowing geographical weighting in model training and prediction as well as allowing hyperparameters, which include both the model hyperparameters and the bandwidth parameter that controls the geographical weighting, to vary spatially. In this regard, both complex spatial and non-spatial processes can be fully considered. We use a random forest model as an example of this generic local modelling framework, which can be naturally extended to other popular models such as neural networks and gradient boosting, for both regression and classification tasks. We benchmark its performance against other global and local statistical and ML alternatives with an empirical flash flood model of Jiangxi Province, China.

2 Methods

Four models are included in comparison to predict a binary flash flood occurrence: 1) logistic regression (LR); 2) geographically weighted logistic regression (GWLR); 3) random forest (RF) and 4) geographically weighted RF (GWRF). They represent the four quadrants of model (as shown in Table 1) types consisting of global/local and statistical/ML, respectively.

■ **Table 1** Four model types.

Model Type	Global	Local
Statistical	Logistic Regression (LR)	GW Logistic Regression (GWLR)
Machine learning	Random Forest (RF)	GW Random Forest (GWRF)

■ **Listing 1** GWRF Algorithm

```

For each location in all locations:
  1. Find a set of hyperparameters and local bandwidth that
     minimises geographically weighted loss with a 5-fold cross
     validation;
  2. Train the local RF model using the best set of
     hyperparameters and local bandwidth;
  3. Use the local RF to predict at any unknown locations weighted
     by its distance away from unknown locations;

Sum of all the distance weighted predictions to be the final
predictions.

```

LR is a global statistical model used to predict binary outcomes. It's a linear model with a logit link function that transforms continuous outcomes into probabilities bounded between 0 and 1. GWLR is a local statistical approach that accounts for location-specific effects when generating the outcome of interest. It fits a geographically weighted logistic regression model at each location using a distance decay kernel governed by a kernel function and kernel bandwidth. This approach allows for parameters in the model to vary spatially. RF is a machine learning algorithm that utilizes ensemble learning methods to make predictions by combining multiple decision trees. While RF is widely used in various applications due to its flexible and accurate predictions, it's considered a global model since the same hyperparameters that govern the tree structure remain constant regardless of geographic location. The last model GWRF is the proposed approach. It trains a separate local RF model at each location allowing different hyperparameters for the RF model and bandwidth for geographical weighting. Each local RF is optimised using a geographically weighted loss function. Then the prediction at an unseen location can be computed as the distance weighted predictions from all RFs. The specific training and prediction process are described as follows:

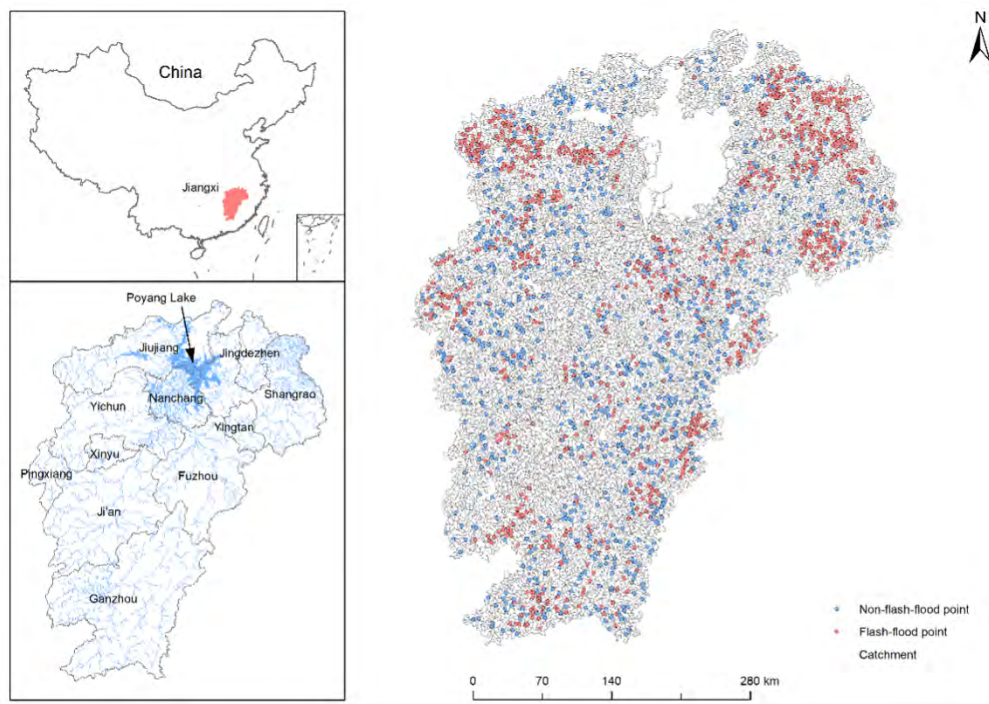
LR and RF are implemented using the sklearn python package [11], GWLR is fitted using the mgwr python package [10], and GWLR is implemented using both sklearn and mgwr. Code and data that produce the results can be found at this repository: https://anonymous.4open.science/r/global_local_ML_GIScience-48F9.

3 A Case Study of Jiangxi Province, China

3.1 Study area

The case study area is Jiangxi, a province in south-eastern China. Jiangxi has long been one of the places suffering flash floods every year in China, which is primarily due to its unique geography and climate. It is located in a mountainous region with over 3,000 rivers and lakes, which accounts for 78% of the total area. The largest freshwater lake in China,

Poyang Lake, is located in the north of the province. Further, Jiangxi is in a subtropical climate zone and experiences a high amount of rainfall during the monsoon season from May to September. Flash flood risk reduction and management is a major challenge to local government with respect to sustainable development. In addition to dams and other flood control infrastructure, mapping flash flood susceptibility has become an effective measure to assist with land use planning as well as to improve public knowledge of flash floods.



■ **Figure 1** Historical flash flood events in Jiangxi Province, China.

3.2 Data

The main dataset used in this research is the flash flood inventory map provided by the Flood Control and Drought Relief Division, Emergency Management Department of Jiangxi, which contains historical flash floods in Jiangxi during 1950-2015. Among 12,388 catchments within the province, 940 contain historical flash flood events. Accordingly, 971 catchments without historical flash floods are randomly selected across space. The final dataset contains 1,911 observations labelled either 1 (flash floods) or 0 (non-flash floods). The resulting flash floods distribution map can be seen in figure 1.

In addition, four ancillary datasets are used to derive potential factors that contribute to flash floods, including the DEM dataset of China (2014), Statistical Parameter Atlas of Rainstorms in China (2010), River System in China (2012) and the Landsat 7 Collection 1 Tier 1 Annual NDVI Composite. Based on those datasets, 10 influencing factors are calculated or extracted: slope, elevation, shape factor, concentration gradient, topographic wetness index, rainfall, peak discharges per unit area, time of concentration, normalized difference vegetation index (NDVI) and distance to the nearest river, which are selected based on previous studies and data availability.

3.3 Results

The dataset was split 80/20, with 20% of the unseen data being used for out-of-sample accuracy assessment, the results of which are shown in Table 2. Three accuracy measures are included:

$$\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{FP} + \text{TN} + \text{FN})$$

$$\text{Recall} = \text{TP} / (\text{TP} + \text{FN})$$

$$\text{Precision} = (\text{TP} / (\text{TP} + \text{FP}))$$

where TP is True Positive; TN is True Negative; FP is False Positive; and FN is False Negative.

■ **Table 2** Accuracy, recall and precision for four models.

Model	Accuracy	Recall	Precision
LR	0.70	0.80	0.66
GWLR	0.75	0.60	0.86
RF	0.81	0.65	0.96
GWRF	0.85	0.76	0.91

Regarding the overall accuracy of models, local models have been observed to have approximately a 5% advantage over their global counterparts. This suggests that allowing parameters to vary spatially can lead to an increase in model accuracy. Furthermore, machine learning (ML) approaches have been found to be approximately 10% more accurate than statistical approaches, indicating that complex non-linear and interaction effects are present and can be captured by ML but not by statistical approaches. The proposed GWRF, which allows for non-linearity, interaction, and spatial heterogeneity, has emerged as the best-performing model, achieving a promising overall accuracy of 85%. Additionally, the GWRF model demonstrates the second-highest precision and recall, resulting in a more well-rounded and balanced performance in estimating flash flood occurrences.



4 Summary

Flash floods can pose significant threats to the environment, properties, and life. Recent advances in AI particularly ML techniques provides new opportunities for assessing and estimating the susceptibility of flash floods – an effective measure that can help with designing flash flood risk reduction strategies. This research develops a novel Geographically Weighted Random Forest (GWRF) within a generalisable local ML framework and compares against other local and global statistical and machine learning approaches in estimating flash flood susceptibility. The preliminary results show that GWRF has the best performance among others with higher accuracy and more balanced precision and recall. The initial findings suggest the importance of incorporating geographic space into ML approaches to improve model performance. However, one drawback of ML is its black-box nature, which limits interpretability. The recent development of eXplainable AI methods (XAI) offers opportunities to estimate the effects of ML models and has been demonstrated to be effective when modeling spatial data [7]. The next step of this research is to investigate the explainability of the ML model to explore spatial and non-spatial relationships, enhancing better understanding of flash flood processes.

References

- 1 Jialei Chen, Guoru Huang, and Wenjie Chen. Towards better flood risk management: Assessing flood risk and investigating the potential mechanism based on machine learning models. *Journal of environmental management*, 293:112810, 2021.
- 2 Romulus Costache, Haoyuan Hong, and Quoc Bao Pham. Comparative assessment of the flash-flood potential within small mountain catchments using bivariate statistics and their novel hybrid integration with machine learning models. *Science of The Total Environment*, 711:134514, 2020.
- 3 Stefanos Georganos, Tais Grippa, Assane Niang Gadiaga, Catherine Linard, Moritz Lennert, Sabine Vanhuyse, Nicholas Mboga, Eléonore Wolff, and Stamatis Kalogirou. Geographical random forests: a spatial extension of the random forest algorithm to address spatial heterogeneity in remote sensing and population modelling. *Geocarto International*, 36(2):121–136, 2021.
- 4 Stefanos Georganos and Stamatis Kalogirou. A forest of forests: A spatially weighted and computationally efficient formulation of geographical random forests. *ISPRS International Journal of Geo-Information*, 11(9):471, 2022.
- 5 Julian Hagenauer and Marco Helbich. A geographically weighted artificial neural network. *International Journal of Geographical Information Science*, 36(2):215–235, 2022.
- 6 Khabat Khosravi, Hamid Reza Pourghasemi, Kamran Chapi, and Masoumeh Bahri. Flash flood susceptibility analysis and its mapping using different bivariate models in iran: a comparison between shannon’s entropy, statistical index, and weighting factor models. *Environmental monitoring and assessment*, 188:1–21, 2016.
- 7 Ziqi Li. Extracting spatial effects from machine learning model using local interpretation method: An example of shap and xgboost. *Computers, Environment and Urban Systems*, 96:101845, 2022.
- 8 Meihong Ma, Changjun Liu, Gang Zhao, Hongjie Xie, Pengfei Jia, Dacheng Wang, Huixiao Wang, and Yang Hong. Flash flood risk analysis based on machine learning techniques in the yunnan province, china. *Remote Sensing*, 11(2):170, 2019.
- 9 Amir Mosavi, Pinar Ozturk, and Kwok-wing Chau. Flood prediction using machine learning models: Literature review. *Water*, 10(11):1536, 2018.
- 10 Taylor M Oshan, Ziqi Li, Wei Kang, Levi J Wolf, and A Stewart Fotheringham. mgwr: A python implementation of multiscale geographically weighted regression for investigating process spatial heterogeneity and scale. *ISPRS International Journal of Geo-Information*, 8(6):269, 2019.
- 11 Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.
- 12 Binh Thai Pham, Mohammadtaghi Avand, Saeid Janizadeh, Tran Van Phong, Nadhir Al-Ansari, Lanh Si Ho, Sumit Das, Hiep Van Le, Ata Amini, Saeid Khosrobeigi Bozchaloei, et al. Gis based hybrid computational approaches for flash flood susceptibility assessment. *Water*, 12(3):683, 2020.
- 13 Seann Reed, John Schaake, and Ziya Zhang. A distributed hydrologic model and threshold frequency-based method for flash flood forecasting at ungauged locations. *Journal of hydrology*, 337(3-4):402–420, 2007.

Understand the Geography of Financial Precarity in England and Wales

Zi Ye¹  

Department of Geography and Planning, University of Liverpool, UK

Alex Singleton  

Department of Geography and Planning, University of Liverpool, UK

Abstract

Financial precarity is a growing and pressing issue in many countries, which refers to a precarious existence which lacks job security, predictability, and psychological or material welfare. Its negative effects can be observed in cognitive functioning, emotional stability and social inclusion. Financial precarity has been proved to be impacted by multifaceted factors ranging from poor quality, unpredictable work, unmanaged debt, insecure asset wealth and insufficient money and resource. However, the geographical variation of financial precarity and the embedded social-spatial inequalities remain understudied. This paper addresses this research gap by introducing a new geodemographic classification of financial precarity, which is developed from a series of small area measurements covering employment, income, asset, liability and lifestyle characteristics of neighbourhoods. The research is conducted within the spatial extent of England and Wales.

2012 ACM Subject Classification Applied computing → Economics

Keywords and phrases Financial precarity, Geodemographic classification, Household finance, Financial Wellbeing

Digital Object Identifier 10.4230/LIPIcs.GIScience.2023.87

Category Short Paper

Funding This work is funded by the UK ESRC Consumer Data Research Centre (CDRC) grant 21 reference ES/L011840/1.

1 Introduction

Precarity has been broadly used to define the state of lacking security and predictability of material and psycho-social deprivation [1]. Particularly, financial precarity refers to the precarious state of being financially insecure or at risk of economic hardship. In social science research, the concept of precarity has been closely associated with employment and work[2][7]. However, since Ettliger first argued for an “unbounded approach” to study precarity[3], there is a growth in geographical understanding of the multi-dimensional nature of contemporary precarity[11]. A main contribution from geographers relates to the spatialisation of precarity, and its situation as a feature of broader life rather than something specific to related to work or income[10]. Such a dualistic characterisation of approach leads to the concept of precarity encompassing both “labour” and “life”, and also lays the foundation for our research to understand the geography of financial precarity. Previous research recognises the detrimental consequences of financial precarity[6] and investigates the structural and institutional drivers of these patterns[5]. There is however dearth of understanding about the geographic variation and characteristics of financial precarity, especially at the small spatial scale.

¹ Corresponding author



© Zi Ye and Alex Singleton;
licensed under Creative Commons License CC-BY 4.0

12th International Conference on Geographic Information Science (GIScience 2023).

Editors: Roger Beecham, Jed A. Long, Dianna Smith, Qunshan Zhao, and Sarah Wise; Article No. 87; pp. 87:1–87:6

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

This research aims to develop new understand about the multidimensional nature and geography of financial precarity through a geodemographic framework encompassing measures of employment status, income and benefits, assets, liabilities and lifestyle factors. The output is a Financial Precarity Classification which maps the residential differentiation of financial precarity across different neighbourhoods in England and Wales at the small area level.

2 Material and methodology

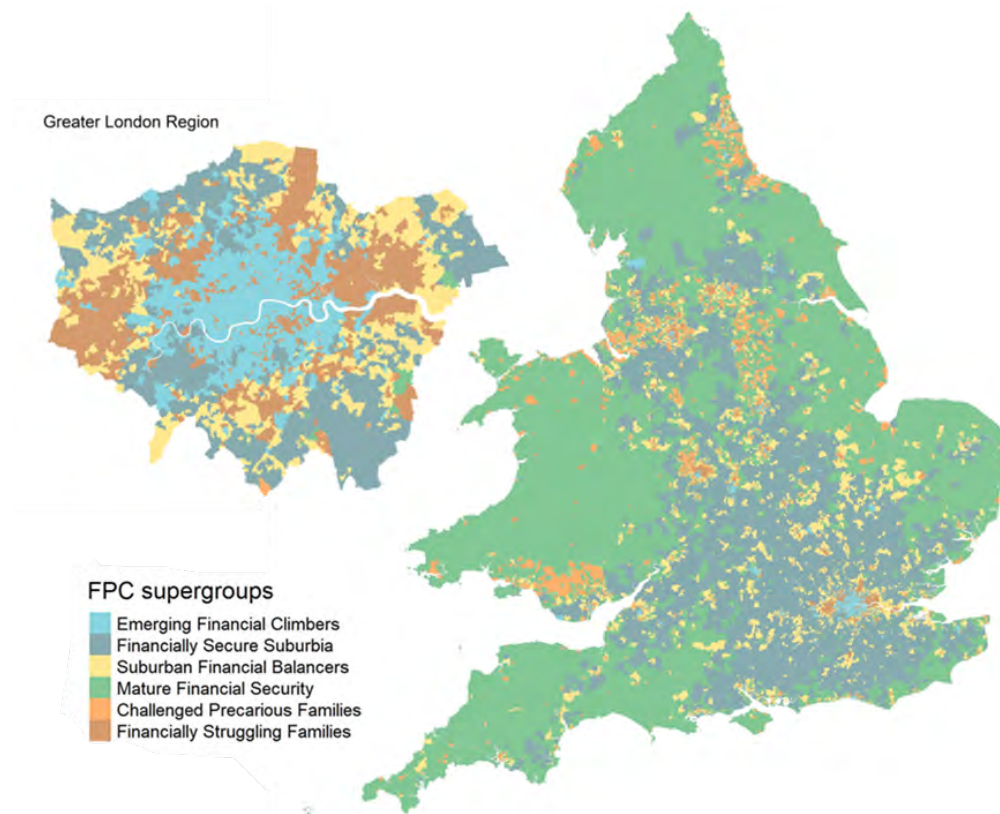
This research captures a variety of structural factors collected at the Lower Super Output Area (LSOA) level to depict the multidimensional facets of financial precarity. These small area measures are used as inputs to a geodemographic classification which groups the 35,672 LSOA zones of England and Wales into different clusters by the common shared salient characteristics. A framework for the typology was developed around five domains related to the main drivers and influences of financial precarity; including “Employment”, “Income”, “Assets”, “Liabilities” and “Lifestyle”. There are further disaggregated into a series of dimensions, which are used as the basis for identifying measures.

Data used to create measures were derived from a variety of sources including the 2021 Census (covering all the employment measures and other dimensions ranging from housing tenure, second address, overcrowding, cars, age band, household composition and health). Other secondary datasets such as Department for Work and Pensions (DWP) Statistics, Energy efficiency statistics, UK Finance Statistics and County Court Judgement (CCJ) Records are used to describe the aggregated level of social benefits, energy consumption, loan and lending, and debt in the LSOAs. In addition, two customer behaviour surveys - GambleAware Treatment and Support Survey and the FCA Financial Behaviour Survey - were included as alternative to derive measures through small area estimation microsimulation. After the correlation analysis between each of the candidate variables, These there are 52 measures formed the input variables to the classification (as listed in Figure 2).

The direct and small area estimated measures offer insights into a spectrum of spatial inequality of financial precarity from multifaceted perspectives. But such multidimensional results are hard to interpret or draw insightful conclusions in isolation. Geodemographics classification is a computational technique used to cluster small areas according to the similarity in area level characteristics[4]. It is a well established and effective method to highlight salient multidimensional characteristics from a body of small area measures. Geodemographic classification has adopted in numerous contexts to create neighbourhood classifications, for example, related to education or digital inclusion[12][8]; and is widely used in consumer segmentation for marketing and other business practices[9].

3 A Classification of the Financial Precarity

A K-means clustering algorithm was implemented to develop the multivariate classification after the standardisation and normalisation of the input variables. The standardisation includes centring and scaling, which transformed the variables to mean zero and standard deviation 1; the normalisation is was conducted through Box Cox transformation which transforms the variables to a normal distribution. Before implementing K-means clustering, a Clustergram was used to decide on the number of clusters of the Supergroup, and after partitioning, the process was repeated for each of the Supergroups to determine the number of Subgroups. As a result, the geodemographic classification clustered the 35,672 LSOA across England and Wales into 6 Supergroups and 14 subgroups. Here we present the characteristics of the 6 Supergroups. Figure 1 shows these on a map for England and Wales, with Greater London as inset.



■ **Figure 1** A map of the Financial Precarity Classification (FPC) for England and Wales.

To better understand the salient characteristics of the classification outcomes, the input variables of each Supergroup were compared with the mean scores for their cluster. These scores are visualized in Figure 2. A significantly higher or lower index score indicates the measure is a key differentiator of the Supergroup. By doing so, each of the 6 Supergroups can be profiled by the combination of their salient measures.

The interpretation of the geodemographic classification is normally based on the index score list as in Figure 2 and presented with labels and descriptions. The “pen portraits” of the Financial Precarity Classification Supergroups are as follows:

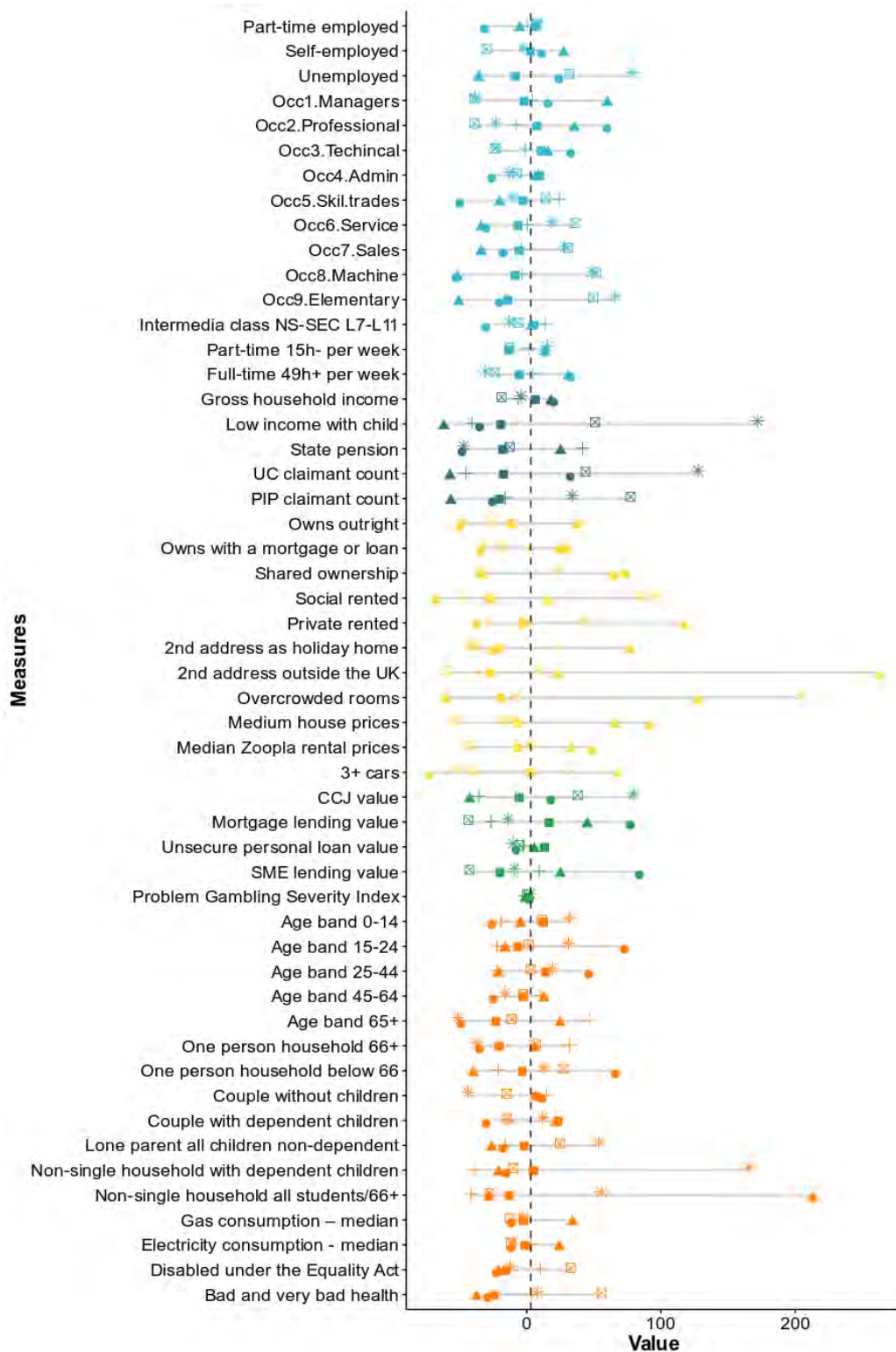
1: Emerging Financial Climbers (8.5%)

Predominantly located in London and other provincial cities and comprises mainly of young professionals and full-time students. This well-educated young Supergroup exhibit the lowest rates of asset ownership such as houses and cars within the UK, and a lack of savings and investment. They typically reside in expensive neighbourhoods and pay high private rentals for over-occupied houses. The younger age of this group is also associated with better health conditions.

2: Financially Secure Suburbia (16.3%)

Residents of these areas mainly consist of family households living as couples, with and without dependent children. A notable characteristic of this group is their financial security, typically with significant financial assets. They not only own houses outright in expensive

87:4 Understand the Geography of Financial Precarity in England and Wales



■ Figure 2 Index scores of the 6 Financial Precarity Classification Supergroups.

areas with high sales and rental prices, but often also own a second address as a holiday home, and possess more than two cars. Their houses are typically under-occupied, leading to the highest gas and electricity consumption among all supergroups. This group also shows a clear inclination towards the elite, particularly in employment as managers, directors, and senior officials.

3: Suburban Financial Balancers (19.2%)

This group has the highest employment rate, which even spread across all spectrums of occupations but higher in the administrative and secretarial occupations compared to other Supergroups. Its secured employment status also embodies in especially low rate in part-time job less than 15 hours. It is a rather average group with few noteworthy variables – only couples with dependent children, partial ownership of houses (own with a mortgage or shared ownership) and personal loans.

4: Mature Financial Security (25.4%)

Residents of these areas are characterised by single-family households consisting of people aged 66 and over. They have solid property assets including outright ownership of houses and multiple cars, although do not necessarily live in the most upscale neighbourhoods. They live comfortably as their houses are more likely to be under-occupied. While retirement and state pensions are common, these areas also observe a mild prevalence of skilled trades occupations. Despite not having the highest gross household income, this group has a quite solid financial status - with an extremely low value for the CCJ debts and high levels of saving and investments. As the area has an ageing population, their health is below the national average.

5: Challenged Precarious Families (20.9%)

This working-class group shows a significant share of social rented housing, with a high incidence of poor health and disabilities, resulting in a notable score in Personal Independence Payment (PIP) benefit claimant counts but more moderate Universal Credit claimants. It also has a relatively lower level of education. Poorer health also leads to economic stress, through higher unemployment and the lowest average household income. Employment tends to be in operational and elementary occupations, routine or semi-routine, service and sales jobs. There are also higher instances of lone parents and dependent children in this group.

6: Financially Struggling Families (9.6%)

This is overall the most financially vulnerable Supergroup, with high levels of unemployment, financial vulnerability, rates of problem gambling, outstanding debts and low income. Given their financial precarity, rates of savings and investments or other property assets like houses and cars are very low. Such issues are exacerbated as there are high instances of dependent children and lone parents, with households often being overcrowded. There are a high proportion of residents below 65 and high rates of unemployment. Those who are working, tend to be in elementary, operational and services occupations. As a result, this group relies heavily on social benefits like Universal Credit and PIP.

4 Discussion and Future Work

Financial precarity is a complex issue, particularly given the current context of the cost of living crisis. In this paper, we developed the Financial Precarity Classification in England and Wales to examine the geographic variation of the socio-spatial inequalities in household financial precarity. A wide spectrum of measures was incorporated to provide unique evidence from separate domains of Employment, Income, Asset, Liability and Lifestyle. The Financial Precarity Classification is a two-tier typology and in future work we will discuss the second tier of the classification. A further agenda is the evaluation of the classification. The internal validation has been conducted with the help of FCA financial behaviour surveys. But external evaluation is also necessary to examine the utility of the classification. The work has the potential to provide spatial insights to policymakers and practitioners associated with the household financial supports and wellbeing. In future, we also plan to consider the temporal changes in the classification to understand the dynamics of neighbourhood financial precarity.

References

- 1 Gabriella Alberti, Ioulia Bessa, Kate Hardy, Vera Trappmann, and Charles Umney. In, against and beyond precarity: Work in insecure times. *Work, Employment and Society*, 32(3):447–457, 2018. Publisher: SAGE Publications Ltd. doi:10.1177/0950017018762088.
- 2 Tom Barnes. Pathways to precarity: Work, financial insecurity and wage dependency among australia’s retrenched auto workers. *Journal of Sociology*, 57(2):443–463, 2021. doi:10.1177/1440783320925151.
- 3 Nancy Ettliger. Precarity unbound. *Alternatives: Global, Local, Political*, 32(3):319–340, 2007. Publisher: SAGE Publications Inc. doi:10.1177/030437540703200303.
- 4 Richard Harris, Peter Sleight, and Richard Webber. *Geodemographics, GIS and Neighbourhood Targeting*. John Wiley & Sons, Ltd, 2005. Publication Title: Geodemographics, GIS and Neighbourhood Targeting ISSN: 1746-0166. doi:10.1057/palgrave.ddmp.4350070.
- 5 Mindaugas Leika and Daniela Marchettini. A generalized framework for the assessment of household financial vulnerability. *IMF Working Paper No. 17/228*, 2017. URL: <https://papers.ssrn.com/abstract=3079554>.
- 6 Jirs Meuris and Carrie Leanaa. The price of financial precarity: Organizational costs of employees’ financial concerns. *Organization Science*, 29(3):398–417, 2018. doi:10.1287/ORS.2017.1187.
- 7 Tajudeen Oluwafem Noibi, Digvijay Pandey, and Adrian Botello Mares. Understanding the concept of the precarity: mirroring colonia mexico 68. *GeoJournal*, 87:5251–5263, 2022. ISBN: 0123456789. doi:10.1007/s10708-021-10562-8.
- 8 Alex Singleton, Alexandros Alexiou, and Rahul Savani. Mapping the geodemographics of digital inequality in great britain: An integration of machine learning into small area estimation. *Computers, Environment and Urban Systems*, 82:101486, 2020. Publisher: Elsevier Ltd. doi:10.1016/j.compenvurbsys.2020.101486.
- 9 Alexander D Singleton and Seth E Spielman. The past , present , and future of geodemographic research in the united states and united kingdom in the united states and united kingdom. *The Professional Geographer*, 0124, 2014. doi:10.1080/00330124.2013.848764.
- 10 Kendra Strauss. Labour geography 1. *Progress in Human Geography*, 42(4):622–630, 2018. doi:10.1177/0309132517717786.
- 11 Louise Waite. A place and space for a critical geography of precarity? *Geography Compass*, 3(1):412–433, 2009. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1749-8198.2008.00184.x>. doi:10.1111/j.1749-8198.2008.00184.x.
- 12 Lili Xiang, John Stillwell, Luke Burns, Alison Heppenstall, and Paul Norman. A geodemographic classification of sub-districts to identify education inequality in central beijing. *Computers, Environment and Urban Systems*, 2018. doi:10.1016/j.compenvurbsys.2018.02.002.

Understanding Active Travel Networks Using GPS Data from an Outdoor Mapping App

Marcus A. Young   

Transportation Research Group, University of Southampton, UK

Abstract

To support a shift to active travel there is a vital need for better data to understand active travel networks: their extent, attributes and current utilisation. Using a big dataset of volunteered geographic information from an outdoor mapping smartphone app, a methodology has been developed to analyse recorded routes to identify missing links in a routable street and path network and to visualise the relative importance of different links of the active travel network. This methodology has then been used to analyse the network for a case study area around Winchester, UK, with new pathways equivalent to 8% of the existing network dataset identified. The automated method developed can be readily applied to other locations and the outputs used to augment existing network datasets and to inform the planning and development of active travel infrastructure.

2012 ACM Subject Classification Applied computing → Transportation

Keywords and phrases active travel, map construction, GPS, volunteered geographic information

Digital Object Identifier 10.4230/LIPIcs.GIScience.2023.88

Category Short Paper

1 Introduction

To meet the UK's target of net zero carbon emissions by 2050, tackle local air pollution, and address the obesity public health crisis, there needs to be a step-change in the use of active travel modes¹. However, the data needed by users of the active travel network, and those who maintain and develop it, is lacking. There is no single dataset that encompasses the entirety of the active travel network, and limited information on its utilisation and attributes. This gap in data provision poses a significant challenge to the successful delivery of active transport policies.

Existing routable network datasets are primarily focussed on meeting the needs of motorised transport and are unsuited to the accurate analysis and planning of active travel.² They mainly consist of street centrelines with attributes that are inadequate for reliable and safe pedestrian and cyclist routing (for example, pavement presence information is lacking). Relatively little attention has been given to how well these data describe the actual active travel network, in terms of both scope and real-world usability. Data from other tracking apps have been used to understand active travel, most notably Strava Metro (for example, see [5] for a review of its use to monitor cycling). However, the Strava dataset is mapped to OpenStreetMap ways so cannot be used to identify unknown parts of the network.

The research described in this paper is part of a larger on-going project - Routable Active Travel Infrastructure Network (RATIN) - being carried out by researchers at the University of Southampton and funded by Ordnance Survey. Phase one of the project was a scoping study to identify data and methods which could help provide a comprehensive routable active

¹ Active travel is defined as journeys made by transport modes that are fully or partially people-powered, irrespective of journey purpose, for example: walking, using wheelchairs, and cycling (including e-bikes).

² The two main options currently available for generating routable transport networks in Great Britain (GB), are OpenStreetMap (OSM) and the MasterMap Highways Network from Ordnance Survey.



© Marcus A. Young;
licensed under Creative Commons License CC-BY 4.0

12th International Conference on Geographic Information Science (GIScience 2023).

Editors: Roger Beecham, Jed A. Long, Dianna Smith, Qunshan Zhao, and Sarah Wise; Article No. 88; pp. 88:1–88:6

Leibniz International Proceedings in Informatics



LIPIC Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

travel network dataset. This included a work package to develop methods for identifying and understanding real world active travel networks from volunteered geographic information in the form of GPS traces derived from smartphone apps. In particular, the research sought to analyse a large dataset from OSMaps.³ The aim of the research described in this paper was to process GPS recorded routes to identify parts of the active travel network that are not currently incorporated in the Ordnance Survey (OS) Mastermap (MM) Highways and Urban Paths products and to assess the suitability of the data to understand where active travel is taking place. The only previous research to use this dataset developed a classification to describe and group walking routes based on environmental characteristics [2].⁴

2 Data and Methods

2.1 Data and study area

The OS smartphone app, OSMaps, is a popular outdoor mapping product in GB that enables users to plan future routes (by manually plotting them) or to record routes (via a smartphone built-in GPS device) as they walk or cycle (or undertake other outdoor activities). Data was provided for all routes intersecting the Hampshire and Southampton Unitary Authority areas that were generated from 2019–2021. Following data cleaning (see section 2.2) and methodological development (see section 2.3), the routes within the Winchester area, NGR squares SU43SE and SU42NE, were analysed.

The raw data (an 80GB CSV file) was imported into a PostgreSQL/PostGIS database and contained 325,532 routes that were entirely within a bounding box around the Hampshire and Southampton boundary. The bulk of the information for each record is held as a complex JSON (not GeoJSON) formatted object in one column. The route information (without GPS timestamps) is stored in a child element that consists of one or more coordinate arrays which each contain two key:value pairs; one for latitude and one for longitude. Native JSON support within PostgreSQL (jsonb) was utilised to extract information from the JSON object making use of lateral joins to generate linestrings from the individual latitude and longitude coordinates making up route features. Information indicating whether the route was “plotted” (recorded by GPS), “routed” (manually created) or “imported” was also extracted.⁵

2.2 Data cleaning

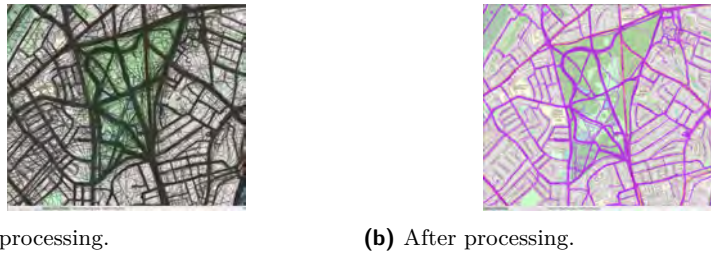
When the extracted routes were visualised it was apparent that the data contained substantial ‘noise’. As shown in Figure 1a, there are many straight lines criss-crossing the area and routes with large distances between vertices making their representation of pathways followed on the ground inadequate. The following inclusion rules were applied using SQL queries to improve the dataset for subsequent analysis: entire route not a straight line; maximum segment length (distance between vertices) < 250 metres; average segment length < 125 metres; and number of segments > 2 (i.e., at least 4 vertices).

The data cleaning process reduced the number of routes from 325,532 to 66,587. The geometry of the cleaned routes is shown for an example area in Figure 1b. The median route length in the cleaned dataset was 6.4km with first and third quartiles of 3.9km and 9.9km respectively. The distribution of routes is left-skewed with a small number of routes

³ See: <https://osmaps.com/en-GB>

⁴ OS has also produced visualisations of the raw data, e.g., see: <https://bit.ly/42HvoY5>.

⁵ A route can consist of a mix of “plotted” and “routed” components.



■ **Figure 1** OSM route data before and after cleaning in an example area.

extending to 75km and beyond. The user specified activity type is overwhelmingly walking (82%), with cycling and running accounting for 10% and 7% respectively, and other activities 1%. The vast majority of the routes are recorded by GPS (62,052 or 93%).

2.3 Map construction methodology

Map construction methods are used to automatically generate (or update) a street network from multiple GPS traces recorded from within vehicles. This is an active research area and many algorithms have been developed to solve the problem (see [1] for a review). In this study, the approach was used to convert the OSM route data into a vector map of the active travel network. A density-based map construction method was used based on the work of [3] and discussed in [1], and implemented using algorithms within the QGIS application. The method consists of line density estimation, skeletonization, conversion to vector data and a topology cleaning/refinement process. The processing steps are detailed below with example outputs shown in Figure 2. Only routes based on GPS were used for the map construction.

Line density estimation

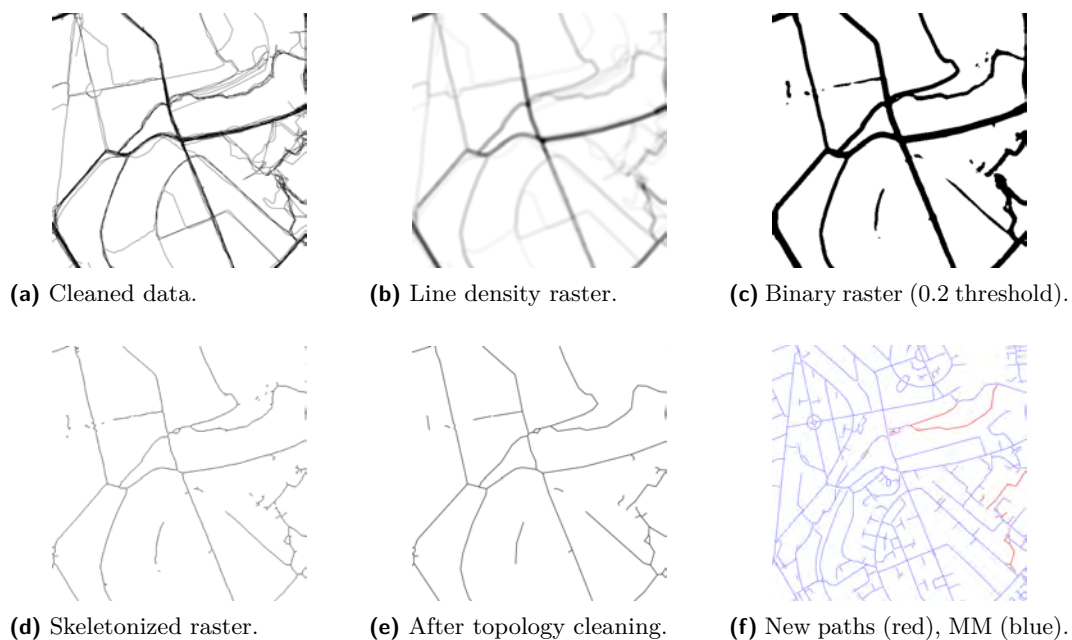
The QGIS line density interpolation tool was used to calculate a density measure of the routes within a circular neighbourhood of each cell across a raster surface. For each raster cell the length of line segments that intersect its circular neighbourhood is summed and divided by the area of the neighbourhood. A cell size of 2m was selected with a neighbourhood radius of 10m. The resultant raster is useful for visualising and analysing levels of use on the active travel network (see Figure 2b). This was followed by conversion to a binary raster, where a threshold line density is applied so that only cells with a high likelihood of a well-used pathway passing through them are assigned a value of 1 (see Figure 2c).

Skeletonization and thinning

The GRASS `r.thin` tool was used to ‘skeletonize’ the binary raster. The tool uses the algorithm described by [4] to thin the non-null cells that represent active travel routes into linear features of single pixel width (see Figure 2d). This step is necessary so that the raster can then be converted into a vector linestring layer using the GRASS `r.to.vect` tool.

Topology cleaning and refinement

The vector layer was simplified by removing vertices with a tolerance of 5m and short dangles up to 15m using the GRASS `v.clean` tool. The resulting vector layer was compared with ground-truth (the known road and urban path network), enabling identification of parts of the active travel map that are not represented by these products. These were extracted



■ **Figure 2** Outputs from the active travel map construction process for a small example area.

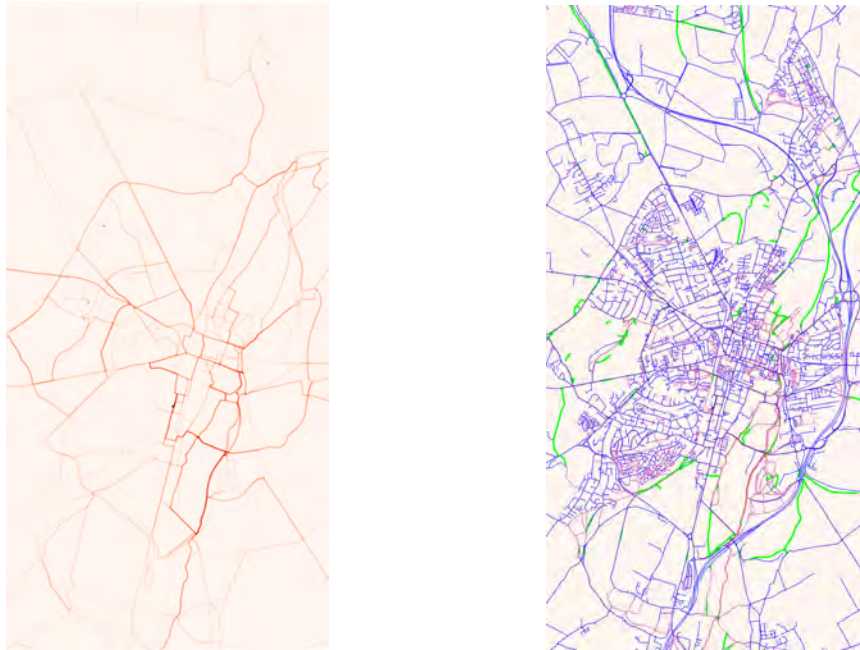
by retaining lines (or line parts) outside of a variable width buffer based on the average real-world width of the MM roadlinks or 5 metres for the urban pathlinks (see Figures 2e and 2f). The PostGIS `ST_ClusterIntersection` function was used to identify interconnected edge clusters which were then converted to simple line features using `ST_CollectionHomogenize`. This enabled the length of these features to be calculated and assessed.

3 Results and Discussion

The line density raster generated for the recorded routes within the two 5km grid squares covering the case study Winchester area is shown in Figure 3a, with darker reds indicating higher line density (and therefore higher level of use). Some of the most popular active travel locations include central streets, such as Christchurch Road, St James' Lane and parts of the High Street, and off-street paths alongside the River Itchen, near Bridge Street and (via Garnier Road) near Saint Catherine's Lock (part of the Itchen Way long distance footpath).

The case study area contains approximately 375km of road links (excluding motorways that are not available for active travel) and 126km of urban path links from the OS MM network dataset. Based on a binary raster with a line density threshold of 1, a total of 38.8km of potential new links was identified, equivalent to 8% of the known network and consisting of 1049 identified line clusters. Many of these clusters are very small artefacts (789 are less than 1 metre) and the median length is 7.7m with the first and third quartiles 1.1m and 22.1m respectively. The top 10% of line clusters are 60m or more in length and account for 28.3km (73%) of the newly identified paths. These are shown in green in Figure 3b overlain on the line density raster with the MM streets and urban paths.

Many of the additional links of the active travel network identified may be available in other datasets, for example the definitive maps of public rights of way (which are available as GIS datasets from some local authorities in GB) or recorded in OpenStreetMap (which may also be partly sourced from the definitive maps, although investigations have shown



(a) LD raster - darker lines indicate greater level of use (more recorded routes). (b) LD raster overlain with MM roads (blue), paths (purple), and new paths >60m (green).

■ **Figure 3** Outputs from the active travel map construction process for Winchester area.

that OSM does not include all the additional links). However, the actual path used may differ from that recorded in the definitive map. An example is shown in Figure 4a, where the extracted path (red) is not recorded as a legal right of way and is the preferred route to the A3090 as indicated by the line density raster (Figure 4b).

Limitations and future work

Users of OSMaps are a self-selecting group of outdoor enthusiasts who have chosen to use this app. It will be biased towards longer leisure journeys rather than shorter travel to work, school or functional journeys. This will affect the relative levels of use of different parts of the network identified by the line density raster. Some important routes, for example a path giving access to a school, may appear insignificant in the line density raster and be excluded when the line density threshold is applied when creating the binary raster. Recorded tracks that on the ground relate to different paths that are close and parallel to one another can appear to be a single path when the binary raster is created. It may be possible to limit this by using a smaller cell size and adjusting the line density neighborhood distance.

Future work will develop methods to automatically link newly identified paths to the existing MM network and seek to improve the extraction of new paths, perhaps using map-matching algorithms that are usually used to identify the road network link (centreline) that is being driven by a vehicle [6]. The use of buffers around paths and roads can be indiscriminate and more problematic in built-up areas where GPS traces can be deflected from their true position by tall buildings. Future work will also consider how this data (potentially with other big data sources) could be incorporated into a decision-making tool that would enable local authorities to understand how active travel networks are being used and thus aid future planning for maintenance, enhancement and additions to infrastructure. This will include further disaggregation of the data to analyse different types of activity.



(a) New path shown in red, existing legal right of way shown as green diamonds. (b) line density raster with line intensity indicating relative level of use of the two paths.

■ **Figure 4** Example of new pathway identified that has higher level of use than existing right of way.

4 Conclusion

While substantial effort is put into monitoring motorized traffic (for example, the statistics compiled by DfT⁶), much less attention, beyond some monitoring of on-street cycle use, is given to understanding the use of active travel networks. This research has shown how a big dataset of volunteered geographic information from an outdoor leisure mapping smartphone app can be used to visualise where active travel is taking place, understand the relative importance of different parts of the active travel network, and through an automated process identify pathways that are not currently contained within a motorized vehicle-oriented street and urban path based network. The automated method that has been developed is readily transferable to other locations and/or other sources of this type of data.

References




- 1 Mahmuda Ahmed, Sophia Karagiorgou, Dieter Pfoser, and Carola Wenk. *Map Construction Algorithms*. Springer, softcover reprint of the original 2015 edition, 2019.
- 2 Andrea Ballatore, Stefano Cavazzi, and Jeremy Morley. The context of outdoor walking: A classification of user-generated routes. *The Geographical Journal*, Advance online publication, 2023.
- 3 James Biagioni and Jakob Eriksson. Inferring Road Maps from Global Positioning System Traces: Survey and Comparative Evaluation. *Transportation Research Record: Journal of the Transportation Research Board*, 2291(1):61–71, 2012.
- 4 B.-K. Jang and R.T. Chin. Analysis of thinning algorithms using mathematical morphology. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(6):541–551, 1990. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.
- 5 Kyuhyun Lee and Ipek Nese Sener. Strava Metro data for bicycle monitoring: a literature review. *Transport Reviews*, 41(1):27–47, 2021.
- 6 Mohammed A. Quddus, Washington Y. Ochieng, and Robert B. Noland. Current map-matching algorithms for transport applications: State-of-the art and future research directions. *Transportation Research Part C: Emerging Technologies*, 15(5):312–328, 2007.

⁶ see: <https://www.gov.uk/government/collections/road-traffic-statistics>

Geography and the Brain's Spatial System

May Yuan   

Geospatial Information Sciences, The University of Texas at Dallas, TX, USA

Kristen Kennedy   

Cognition and Neuroscience, The University of Texas at Dallas, TX, USA

Abstract

Extensive research in spatial cognition and mobility has advanced our knowledge about the effects of geographic settings on human behaviors. This study, however, takes an alternative perspective to examine how the brain's spatial system may mediate the geographic effects on spatial behaviors. Our previous research using data from OpenStreetMap, SafeGraph POIs, and human participants from the National Alzheimer's Coordinating Center (NACC) resulted in a model with 83.33% prediction accuracy from geographic settings to the zonal categorization of the cognitive state based on NACC participants. A follow-up study showed that the complexity of a geographic setting has a direct effect on cortical thickness in the brain's spatial cell system. In this study, we leverage findings from the two studies and interrogate the geographic settings to discern environmental correlates to zonal cognitive categorization. We conclude with thoughts on the implications of brain-inspired GIScience.

2012 ACM Subject Classification Applied computing → Psychology; Applied computing → Health informatics

Keywords and phrases Brain, geographic complexity, mild cognitive impairment, Alzheimer's Disease

Digital Object Identifier 10.4230/LIPICs.GIScience.2023.89

Category Short Paper

Funding This research study is supported in part by (US) National Institute of Health (NIH) grant R21 AG069267 to both authors. In addition, the materials is based upon work supported by (while Yuan is serving at) (US) National Science Foundation (NSF). The views expressed here are the authors' and do not necessarily reflect the views of NIH or NSF. The NACC database is funded by NIA/NIH Grant U24 AG072122 and many grants to various principal investigators.

1 Introduction

Tolman's [13] concept of cognitive maps, as an essential mental representation of space, prevails in GIScience literature. The discoveries of place cells [8] and grid cells [7] in the hippocampal formation in mammalian brains gave new insights into cognitive maps with neural connections and activities. Yet recent advances in neuroscience unveiled the brain's spatial system much different from the conventional GIS. Although what works in the brain may not be the most effective strategy for computer systems, understanding the brain's spatial system may provoke new ideas for spatial encoding or algorithms that are more flexible and perhaps more powerful than what the prevalent GIScience research can offer.

This study is part of a larger project that investigates the correlative effects of environmental complexity on Mild Cognitive Impairment (MCI) and Alzheimer's Disease (AD). The project is based on the premise that people living in a geographically more complex environment are more often able to retrieve information on spatial relations among landmarks and places when navigating the environment. Traffic dynamics further motivate them to build cognitive maps and recall route options. MCI and AD diseases weaken such cognitive mapping abilities as four out of 10 early warning signs of dementia relate to spatial functions (Figure 1).



© May Yuan and Kristen Kennedy;
licensed under Creative Commons License CC-BY 4.0

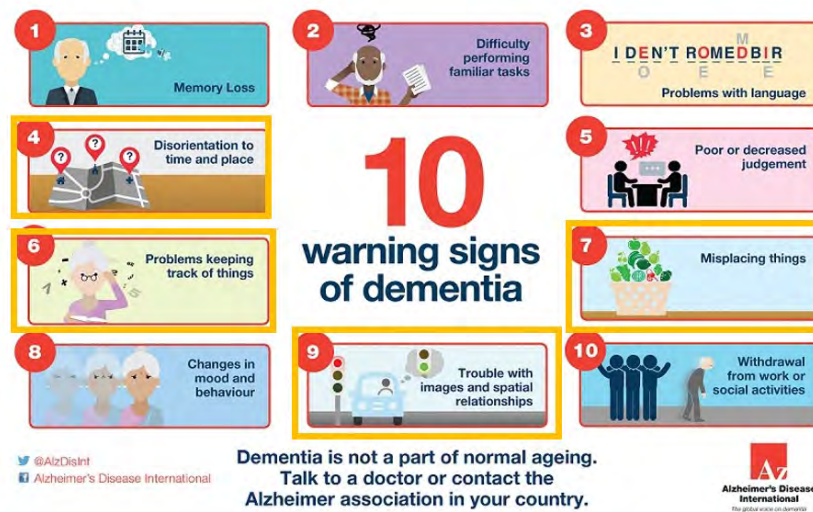
12th International Conference on Geographic Information Science (GIScience 2023).

Editors: Roger Beecham, Jed A. Long, Dianna Smith, Qunshan Zhao, and Sarah Wise; Article No. 89; pp. 89:1–89:7

Leibniz International Proceedings in Informatics



LIPIC Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany



■ **Figure 1** Four out of 10 early warning signs of dementia are related to spatial cognition: 4, 6, 7, and 9.

Moreover, neural research showed early neuropathology of AD in the brain's spatial system [11], leading to spatial navigation impairments which differentiated MCI and AD patients from healthy aging adults [9, 1]. We hypothesize that regularly navigating a complex environment can strengthen the brain's spatial system responsible for cognitive map building and lead to non-pharmaceutical mitigation of MCI and AD.

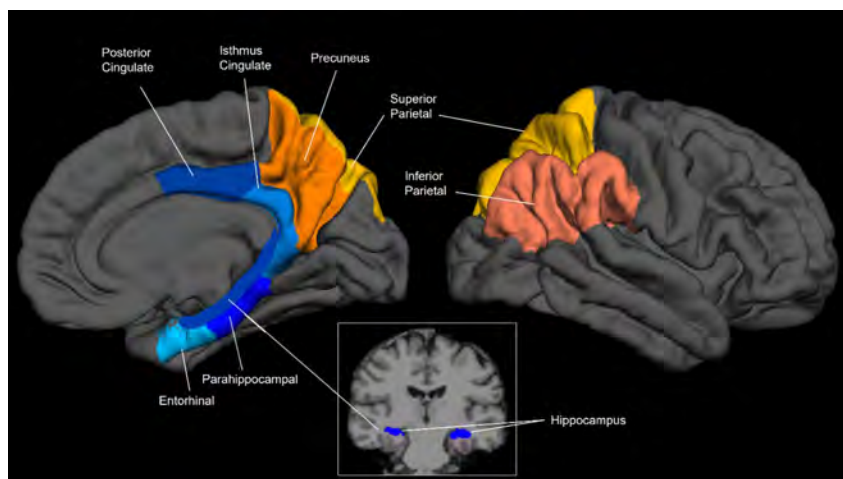
The next section highlights key ideas and findings from the two previous studies as background information. We will then report findings from this study on environmental measures that correlate to zonal cognitive categorization. From the findings and recent advances in the brain's spatial system, we contrast the brain's spatial system with GIScience approaches to spatial representation and computing for potential new ideas moving forward in GIScience.

2 Our recent studies

Lynch's seminal work: *The Image of the City*, defined the concept of city legibility by the pattern of interrelations among five elements: paths, edges, districts, nodes, and landmarks [6]. Our previous study followed Lynch's ideas to compute network measures and points of interest (POI) to represent the complexity of an environment, with which we developed a neural network model to predict zonal cognitive status based on NACC participants' cognitive tests and diagnoses across the US [14]. Both environmental and cognitive measures were summarized into 3-digit zipcode zones, the finest spatial resolution available to researchers. We categorized the cognitive status (normal or AD-inclined) for 154 individual zipcode zones based on cognitive diagnoses (normal, MCI, or AD) of 22,553 NACC participants. Taking the approach of categorical prediction commonly used in medical science, we developed a neural network model that used environmental measures (discussed in Section 3) to predict the cognitive status of each zipcode zone. The input data inherited high spatial heterogeneity and spatial uncertainty. These 3-digit zipcode zones varied from less than 2 to more than 35,000 km^2 . More often than not, environmental measures and cognitive diagnoses were unevenly distributed within individual zones. Participants might travel across zipcode zones or relocate

across zones. Other researchers showed PM2.5, ozone, nitrogen dioxide and nanoparticles and other environmental factors might increase the risk of MCI and AD [5, 4, 10]. Despite the massively noisy data, the model was able to make predictions at 83.87% accuracy, 95.23% precision, 83.33% recall, and 0.89 F1-score. The model suggested AD-inclined zones likely associated with longer street segments, higher circuitry, and fewer points of interest (i.e., lower environmental complexity).

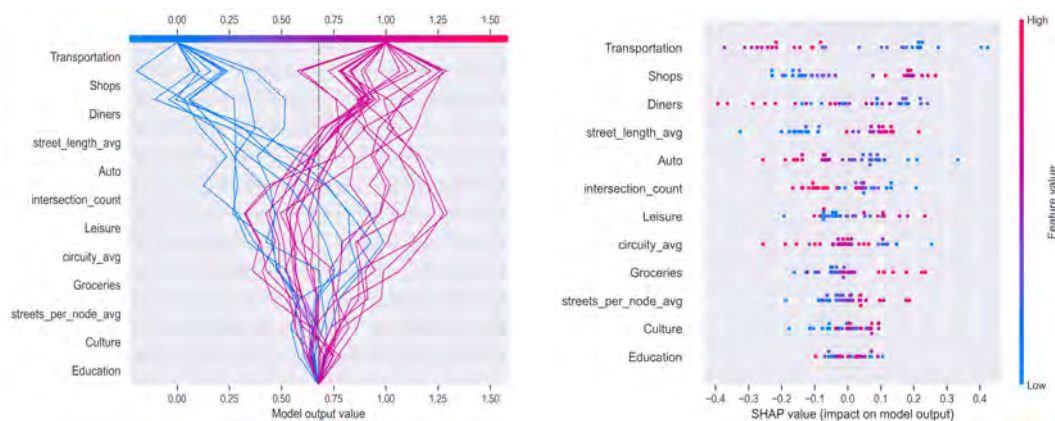
Following the initial study, Shin (2023) explored the associations among environmental complexity, regional brain volumes and cortical thicknesses, against diagnoses of 660 NACC participants with structural brain MRI images[12]. The study compared two sets of brain regions (Figure 2): (1) the hippocampus and the parahippocampal cortex, and posterior cingulate cortex responsible for the allocentric frame of reference in which locations of entities and their relations are external to and independent of the agent who interacts with the environment; (2) the posterior parietal cortex, responsible for the egocentric frame of reference in which all entities, their locations, and relations are based on the agent's location and perspective. ANOVA analyses suggested no interactions between environmental complexity and age on MCI/AD diagnoses, while both showed significant associations. Shin then applied structural equation modeling (SEM) to test the effects of environmental complexity and age on AD diagnoses and the brain's egocentric and allocentric regions and spatial cognition. His SEM suggested a significant effect of higher environmental complexity on a greater volume in the brain's allocentric regions, but not in the egocentric regions. The SEM also suggested a significant pathway with higher environmental complexity to higher allocentric volume then to lower MCI/AD diagnosis. However, the direct effect of environmental complexity on MCI/AD was insignificant. Shin concluded that the relationship between environmental complexity and spatial cognitive deficits in MCI/AD was indirect and was mediated by the brain's allocentric regions. Compared to other social and economic determinants (like gender, income, and education), the direct association of environmental complexity and the allocentric brain implied possibilities of geographically induced neural plasticity and a new role for geography in non-pharmaceutical interventions to MCI and AD.



■ **Figure 2** Distinct brain regions responsible for allocentric (blue shades) and egocentric (orange shade) frames of reference. Adapted from [12].

3 Environmental measures, zonal cognitive prediction, and the brain's allocentric system

Initially, we considered 40 environmental measures from street networks, POI types, and POI distributions. We reduced the number of environmental measures to 12 by removing highly correlated measures and those with extremely skewed distributions across zipcode zones. We applied Shapley additive explanations (SHAP) tools to evaluate the contributions of these environmental measures to model prediction on the test data (Figure 3). The model predicted binary zonal categories: 0 for cognitively-normal zones and 1 for AD-inclined zones. All POI measures (transportation, shops, dining, auto, leisure, groceries, culture, and education) and intersection counts were linear density measures (i.e., frequency over total street length in a zone). Education institutions, cultural landmarks, and averaged number of streets per node (`streets_per_node_avg`) contributed minimally to the model prediction, while intersection counts, auto services, averaged street length, diners, shops, and transportation stations appeared as major discriminators to differentiate normal from AD-inclined zones. Moreover, more transportation stations, diners, auto services, and intersection counts as well as shorter averaged street lengths appeared as the primary push for “normal” predictions (Figure 3b). The other environmental measures showed large overlaps between measures and model prediction and gave mixed signals on their effects on model prediction.



(a) Contributions of environmental measures to model prediction.

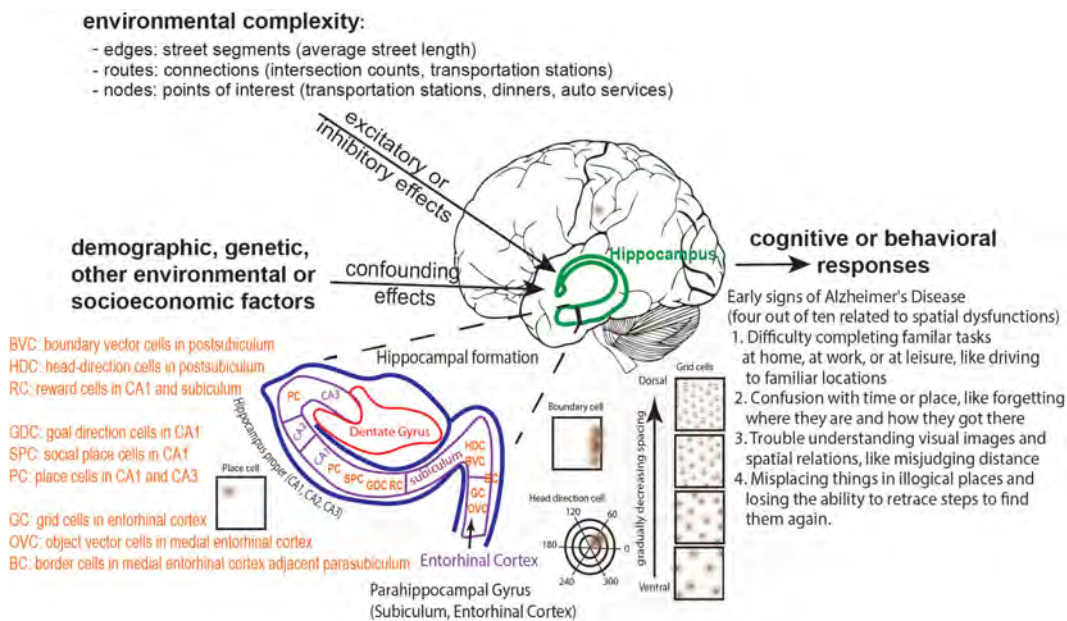
(b) Correlations of model prediction with environmental measures.

■ **Figure 3** Twelve environmental correlates for zonal categorical predictions.

The measure of shop density over street length was counter-intuitive as higher shop density was more correlated to “AD-inclined” prediction (Figure 3b). We consider two possible explanations. First, zoning regulations restrict shops clustered in malls or plazas in the US. Therefore, shops are seldom evenly distributed along a street, so the shop density over unit street length is unlikely meaningful in many US places. Second, except for major stores, such as the Home Depot, Macy’s, or Barnes & Noble, shop signage is often invisible from the street. Therefore, most shops likely provide little help for spatial cognition but may create complications in learning the environment.

The primary environmental measures (transportation, diners, average street length, auto services, and intersection counts) relate well with the brain’s spatial system for allocentric navigation: grid cells, place cells, head orientation cells, boundary cells, object vector cells, and goal direction cells (Figure 4). Grid cells reside primarily in the entorhinal cortex. Each grid cell is responsive to locations in a hexagonal configuration. Grid cells located

towards the dorsal end of the entorhinal cortex are responsive to finer hexagonal grids. Head direction cells, boundary vector cells (a.k.a border cells), object vector cells, reward cells, goal direction cells, place cells, and social place cells in the hippocampal formation have specialized firing fields when the agent (i.e., animals or humans) faces a particular direction, nears a boundary, observes objects, seeks a goal, and recognizes a reward location, one's own location, or locations of one's own kind. Grid cells provide the allocentric reference frame necessary to position objects, boundaries, destinations, one's own location, and others' locations in a common framework and create a cognitive map. The major environmental measures contributed to our prediction model correspond to edges, connections, and nodes in creating navigation routes. Nodes likely correspond to object vector cells, connections to head direction cells, and edges to boundary cells. As one navigates in a geographic environment (i.e., the entire environment is not visible from a single vantage point, and learning the environment requires one to traverse through the environment and mentally integrate spatial observations and experiences from location to location), the sequential firings of place cells and head direction cells, as well as other spatial cells, in the common framework provided by the grid cells allow the hippocampus to construct a cognitive map and perform path integration. Degradation in the entorhinal cortex and hippocampus leads to spatial dysfunctions commonly observed in early MCI and AD patients [1, 3].



■ **Figure 4** Environmental complexity, the brain's spatial system, and spatial cognitive degradation.

There are many neural implications for GIScience. We highlight three points here. First, the multiple resolutions of grid cells collectively fire to transmit signals to downstream spatial cells. Simultaneously imposing hexagon configurations of varying resolutions allows for capturing the bigger picture and fine details for everything, everywhere, all at once. On the contrary, GIS data or functions commonly stay in one scale or resolution. A common practice is to separate data at different resolutions into separate layers or sets. Vertical integration of data representing different themes at different resolutions remains underdeveloped.

Secondly, a place cell's firing field is context-dependent. As one moves from one environment to another, place cells remap firing fields accordingly. Only sparse place cells fire in a given environment. Among the place cells that fire, some place cells fire immediately, but others fire late. The fast-firing place cells are generalists, rapidly recognizing the general spatial configuration of one's location; for example, I am in a school. The late firing place cells refine location recognition to, say, my daughter's high school. By doing so, the brain's allocentric system allows one to recognize the kind of environment quickly and then the specifics of the environment. On the other hand, grid cells have no remap functions, hence providing persistent references to different environments and facilitating spatial integration across environments. Research on geospatial ontology and semantic knowledge graphs has been building hierarchical structures of geographic kinds. Multiple place cells with different responses may give rise to algorithms for ontological or semantic computing.

Thirdly, the current GPS design gives users turn-by-turn instructions or has users follow a blue dot in close view without any geographic context. Even when we safely arrive at a destination on time, we have no idea about where we are and what we have passed by. Like people losing arithmetic skills due to over-reliance on calculators, the popularity of GPS navigation systems likely deskills people's spatial cognition and wayfinding, or worse yet, increases the risk of MCI and AD. Redesigning GPS navigation systems should attend to means that can encourage cognitive map building and attend to geographic contexts. New auditory GPS, for example, promises a viable alternative [2]

The brain encodes and processes spatial information differently from conventional GIS technologies. Many AI researchers seek inspiration from neuroscience to develop new algorithm architectures or learning pipelines. GIScience researchers should also explore the brain's spatial functions, not only for GeoAI but for Brain-inspired GIScience.

References

- 1 Jiri Cerman, Anđel Ross, Jan Laco, Vyhnaek Martin, Nedelska Zuzana, Mokrisova Ivana, Sheardova Katerina, and Hort Jakub. Subjective spatial navigation complaints—a frequent symptom reported by patients with subjective cognitive decline, mild cognitive impairment and alzheimer's disease. *Current Alzheimer Research*, 15(3):219–228, 2018.
- 2 Gregory D Clemenson, Antonella Maselli, Alexander J Fiannaca, Amos Miller, and Mar Gonzalez-Franco. Rethinking gps navigation: creating cognitive maps through auditory clues. *Scientific reports*, 11(1):1–10, 2021.
- 3 DP Devanand, G Pradhaban, X Liu, A Khandji, S De Santi, S Segal, H Rusinek, GH Pelton, LS Honig, R Mayeux, et al. Hippocampal and entorhinal atrophy in mild cognitive impairment: prediction of alzheimer disease. *Neurology*, 68(11):828–836, 2007.
- 4 Carles Falcón, Mireia Gascon, José Luis Molinuevo, Grégory Operto, Marta Cirach, Xavier Gotsens, Karine Fauria, Eider M Arenaza-Urquijo, Jesús Pujol, Jordi Sunyer, et al. Brain correlates of urban environmental exposures in cognitively unimpaired individuals at increased risk for alzheimer's disease: A study on barcelona's population. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, 13(1):e12205, 2021.
- 5 Caleb E Finch and Alexander M Kulminski. The alzheimer's disease exposome. *Alzheimer's & Dementia*, 15(9):1123–1132, 2019.
- 6 Kevin Lynch. *The image of the city*. MIT press, 1964.
- 7 Edvard I Moser, Emilio Kropff, and May-Britt Moser. Place cells, grid cells, and the brain's spatial representation system. *Annu. Rev. Neurosci.*, 31:69–89, 2008.
- 8 John O'Keefe and Jonathan Dostrovsky. The hippocampus as a spatial map: preliminary evidence from unit activity in the freely-moving rat. *Brain research*, 1971.

- 9 Martina Parizkova, Ondrej Lerch, Scott Douglas Moffat, Ross Andel, Adela Fendrych Mazancova, Zuzana Nedelska, Martin Vyhnalek, Jakub Hort, and Jan Laczó. The effect of alzheimer's disease on spatial navigation strategies. *Neurobiology of Aging*, 64:107–115, 2018.
- 10 Ruth Peters, Nicole Ee, Jean Peters, Andrew Booth, Ian Mudway, and Kaarin J Anstey. Air pollution and dementia: a systematic review. *Journal of Alzheimer's Disease*, 70(s1):S145–S163, 2019.
- 11 Alberto Serrano-Pozo, Matthew P Frosch, Eliezer Masliah, and Bradley T Hyman. Neuro-pathological alterations in alzheimer disease. *Cold Spring Harbor perspectives in medicine*, 1(1):a006189, 2011.
- 12 Naewoo Shin. Investigating associations among geospatial environmental complexity, brain morphometry, and cognition across the alzheimer's disease spectrum. Master's thesis, Department of Cognition and Neuroscience, The University of Texas at Dallas, Richardson, Texas, USA, 2023.
- 13 Edward C Tolman. Cognitive maps in rats and men. *Psychological review*, 55(4):189, 1948.
- 14 May Yuan and Kristen M Kennedy. Utility of environmental complexity as a predictor of alzheimer's disease diagnosis: A big-data machine learning approach. *The Journal of Prevention of Alzheimer's Disease*, 10(2):223–235, 2023.

Visual Methods for Representing Flow Space with Vector Fields

Han Zhang ✉

School of Urban Planning and Design, Peking University Shenzhen Graduate School, Shenzhen, Guangdong, China

Key Laboratory of Earth Surface System and Human-Earth Relations of Ministry of Natural Resources of China, Peking University Shenzhen Graduate School, Shenzhen, Guangdong, China

Zhaoya Gong¹ ✉

School of Urban Planning and Design, Peking University Shenzhen Graduate School, Shenzhen, Guangdong, China

Key Laboratory of Earth Surface System and Human-Earth Relations of Ministry of Natural Resources of China, Peking University Shenzhen Graduate School, Shenzhen, Guangdong, China

Jean-Claude Thill ✉

Department of Geography and Earth Sciences, University of North Carolina at Charlotte, NC, USA

Abstract

The issue of human mobility has been a focal point of research among numerous scholars in the field of geography for decades. Among them, the visualization of origin-destination (OD) data is an important branch of geographic flow studies. In this paper, we vectorize and represent immigration flows using OD flow data of U.S. immigrants in the year 2000, constructing an immigration space. Through data validation, it is confirmed that the vector field satisfies the Gauss's theorem and is irrotational, demonstrating that the field can be derived from a potential and that the field is uniquely determined by the potential. Scalar potential fields are inferred from the vector field, providing a more intuitive and convenient description of the underlying flow patterns in geographical interaction matrices. Additionally, this paper employs potential fields and applies a density-equalizing areal cartogram to reconstruct the global representation of functional space, constructing cartogram maps based on potential magnitudes.

2012 ACM Subject Classification Applied computing → Cartography

Keywords and phrases interstate migration, vector field, areal cartogram, geographic visualization

Digital Object Identifier 10.4230/LIPIcs.GIScience.2023.90

Category Short Paper

1 Introduction

The issue of human mobility has long been a focus of research in the field of geography. These studies encompass various areas such as urban spatial structure, urban and regional development, transportation and infrastructure planning, environmental pollution [1], elections and political polarization [3], among others. The visualization of OD data is an important branch of geographic flow research. Plane used a reverse doubly constrained gravity model to calibrate and estimate the cognitive or functional distances between states based on observed interstate migration flows, choosing distances to represent observed flows and visualizing “migration space” as distorted maps [6]. In 1976, Tobler introduced the concept of vector fields and proposed a vector representation method based on OD data, considering the scalar potential of vector fields as a way to describe hidden forces [7]. In recent years, scholars have

¹ Corresponding author



applied vector fields to the study of urban spatial structure. For example, Mazzoli et al. used data from multiple cities to demonstrate that vector fields constructed from flow data satisfy the Gauss (divergence)'s theorem and possess irrotationality. They also explored the utility performance of gravity and radiation models in vector fields based on flow data [4]. Yang, H et al. based on Mazzoli's vector representation method, defined anomalous fields, source fields, and dispersion fields to identify abnormal human flows [9]. Furthermore, scholars have used spatio-temporal potential fields to predict traffic flows [8] and analyzed trade flows in regional science and spatial economics using vector gradient methods and gravity models [5].

Tobler proposed that M_{ij} refers to the flow from location i to location j . Flows occur between various scales, such as communities, cities, and regions. The OD matrix M_{ij} only contains information about the origin and destination of trips and does not include information about intermediate points along the trajectories or visits. The directional attribute of the vector representing the flow is represented by a unit vector from location i to location j . Tobler demonstrated through a series of algebraic transformations that the frequency of the differences between interactions in both directions divided by their sum (i.e. $(M_{ij} - M_{ji}) / (M_{ij} + M_{ji})$) is considerably high and robust. Therefore, this quantity is introduced as a component for constructing vectors. Finally, the vectors pointing from the origin location i to the destination location j are summed, defining a vector field in space. Another approach to construct vectors is directly using net migration flow as the numerator in the vector representation [4]. Additionally, besides the aforementioned methods of constructing vectors, error terms can also be utilized [2]. The vector representation method based on error terms incorporates detailed errors related to observed flows between regions (such as model errors and missing variables like individual preferences), measurement errors (pertaining to the variables themselves), and pure random effects. On the other hand, the vector \vec{c}_i represents half the difference between opposing direction error terms.

Every vector field can be written as the gradient of a scalar field plus an additional vector field. These two components are respectively referred to as the scalar potential and the vector potential. If the second field is everywhere zero and only then, the original vector field can be identified as the gradient of a scalar field. In order to recover this scalar potential, the gradient operation can be reversed through integration to compute the scalar potential. Therefore, if we want to determine the scalar potential of a vector field, it is necessary to ensure that the rotational vector at every point of this vector field is zero, indicating a curl-free condition. The curl-free property of a field implies that the field can be derived from a potential, where the field is uniquely determined by the potential alone.

Currently, research related to representing flows using vectors is relatively scarce. Besides, the majority of existing flow visualization techniques primarily involve flow mapping, which provides a descriptive representation of the flow. In contrast, our method employs the concept of potentials to interpret the flow, offering a distinct visualization approach. As a result, these two methods are not directly comparable. In this study, based on interstate migration data from the United States, we will investigate the problem starting from the core concept of vector space. We will begin with the vector field constructed from OD flow data and generate a potential field through integration.

2 Data and Methods

2.1 Data

The primary data for this study consists of population migration data between states in the United States from the 1965–1970, 1975–1980, 1985–1990, and 1995–2000 censuses. Additionally, data on the physical distances between states during each time period and the population centroid coordinates for each state are included.

2.2 Theoretical models

In terms of vector definition, two different approaches have been identified based on methods applied in various literature. The first approach, proposed by Tobler, involves constructing vectors based on relative net flow. The second approach, employed by Mazzoli, utilizes absolute flow.

(1) Vector Representation Based on Relative Migration Flow

As mentioned earlier, d_{ij} represents the physical distance between the population centroids of locations i and j ; \vec{C}_i denotes the vector aggregation centered at the population centroid of location i , originating from the origin point O . The vector space should contain 48 vectors located at the population centroids of each state, with distinct directions and magnitudes. The formulas are shown in (1) and (2).

$$d_{ij}^2 = (X_j - X_i)^2 + (Y_j - Y_i)^2 \quad (1)$$

$$\vec{C}_i = \frac{1}{n-1} \sum_{j=1}^n \frac{M_{ij} - M_{ji}}{M_{ij} + M_{ji}} \frac{1}{d_{ij}} [(X_j - X_i), (Y_j - Y_i)] \quad (2)$$

(2) Vector Representation Based on Absolute Migration Flow

T_{ij} represents the migration flow from location i to j , \vec{u}_{ij} represents the unit vector (directional attribute) from i to j . Then, the vectors pointing to all destination locations j are summed to obtain the resulting vector at each location i . m_i represents the total outflow from location i . Finally, the vector \vec{W}_i can be constructed as shown in equation (3). These vectors define a field in space, determining the average outward direction of movement at each point.

$$\vec{W}_i = \frac{\vec{T}_i}{m_i} = \sum_j^n \frac{T_{ij}}{m_i} \vec{u}_{ij} \quad (3)$$

$$m_i = \sum_j^n T_{ij} \quad (4)$$

3 Results

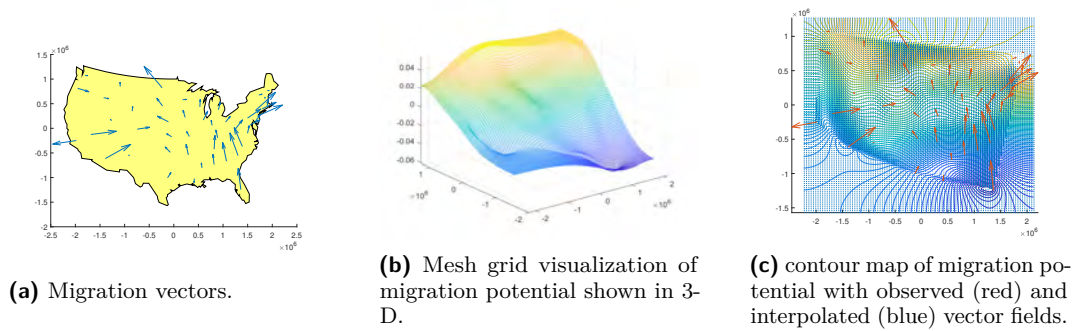
Using US migration data from the period of 1995–2000, we constructed vector fields, three-dimensional grid visualizations of migration potential, and contour maps of migration potential based on the different vector representation methods (Figure 1 and 2).

Among them, the visualizations of the vector representation method based on net flow (Tobler's method) for the time periods of 1965–1970, 1975–1980, 1985–1990 are shown in the following figures (Figure 3, 4, and 5):

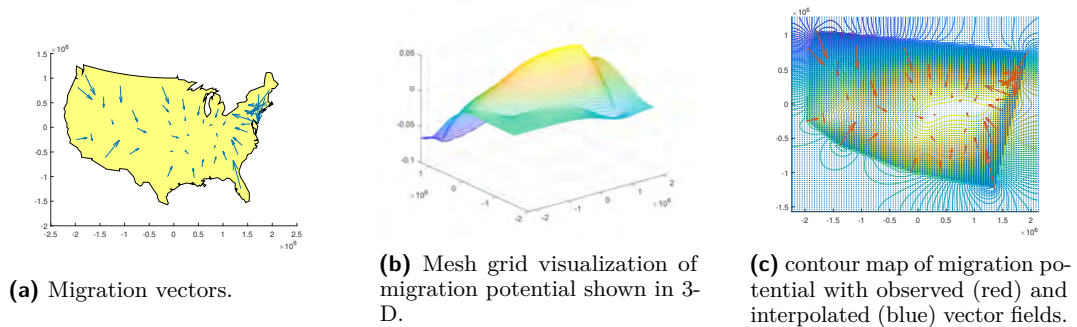
4 Discussion

Based on the migration flows in and out of each state, California consistently ranks at the top in terms of both incoming and outgoing population numbers among all states. Following closely are states like Texas, Florida, and New York, which also exhibit significant flows of population in and out. It is noteworthy that Texas and Florida are located in the moderate climate region known as the "Sun Belt". This aligns with the notable trend of population

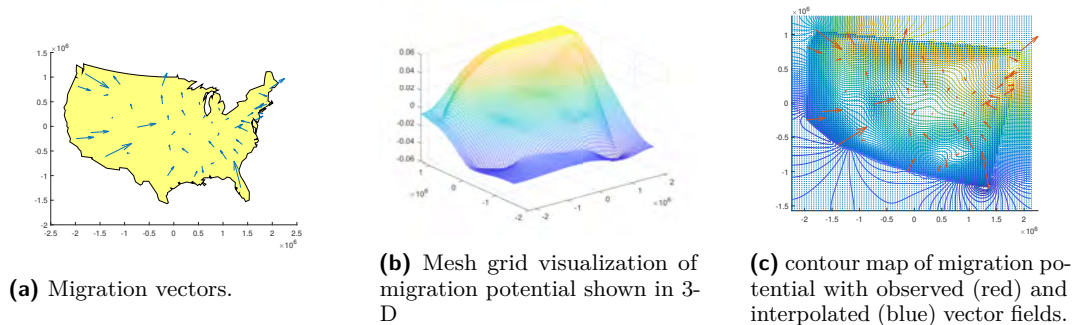
90:4 Visual Methods for Representing Flow Space with Vector Fields



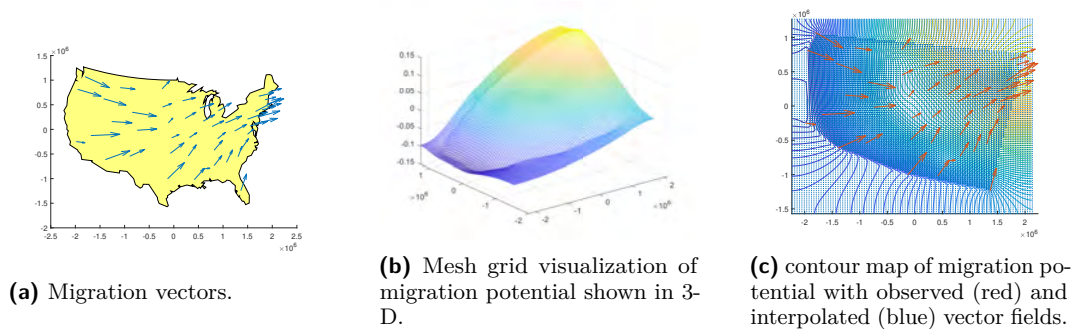
■ **Figure 1** Vector representation based on relative net flow(1995–2000).



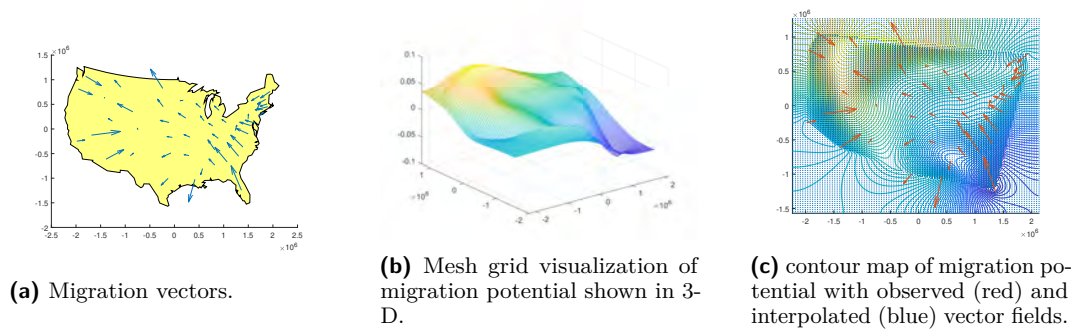
■ **Figure 2** Vector representation based on absolute traffic(1995–2000).



■ **Figure 3** Vector representation based on relative net flow(1965–1970).



■ **Figure 4** Vector representation based on relative net traffic(1975–1980).



■ **Figure 5** Vector representation based on relative net flow(1985–1990).

migration from northern regions of the United States to warmer southern areas, often referred to as the “Sun Belt.” During the time period of 1985–1990, there was a significant decline in both incoming and outgoing population flows across states. This could be attributed to various socio-economic and political factors, including the stabilization of the U.S. economy during that period. Such factors likely influenced people’s decisions regarding migration, leading to a decrease in population mobility between states.

The results obtained from the vector representation method based on absolute flow exhibit overall similarities to those obtained from Tobler’s method, depicting similar patterns at a macro level. This discrepancy arises from the different definitions of vector flows, leading to distinct underlying interpretations of the vector fields. The three-dimensional grid visualization of the scalar potential in Figure 1b) exhibits greater complexity compared to Figures 2b), featuring two peaks. This complexity is particularly meaningful in exploring the implicit forces underlying interactions. Similar to how Tobler conceptualizes the flow field as “wind,” this wind implies a potential function that facilitates interactions in specific directions, aiding in uncovering the causes of asymmetric interactions.

In the vector representation method based on relative net flow, the visualizations in Figures 3, 4, and 5 reveal certain migration patterns during the 1970s and around that time, indicating that the central region of the United States was a primary destination for population movements (Figure 3a)). From the mid to late 1980s, there were noticeable fluctuations in both the origins and destinations of migrants, gradually shifting towards the northeastern part of the country (Figure 4a)). The trend of migration towards the central region persisted in the 1990s (Figure 5a)). However, in the period from 1995 to 2000, significant changes in migration patterns occurred, particularly in California, where a major shift in immigration patterns was observed, along with slight outward migration from some northeastern states (Figure 3a)). Figure 3b), 4b), and 5b) represents the migration potential, revealing implicit forces that can be further explored.

5 Conclusion

Functional spaces are closely connected to human perception and utilization of physical spaces. Conceptualizing the spatial patterns of functional relationships embedded within physical spaces is crucial for understanding the spatial interaction processes that shape geographical phenomena. The use of vector fields to introduce and describe implicit forces in interactions proves valuable. Vector fields approximate the gradient of scalar potentials, which can be used to explain flows. While vector field methods do not directly transform relative distances into functional spaces, they offer an alternative perspective for integrating spatial interaction patterns and aid in developing a global view of functional spaces.

From a visualization perspective of geographic flows, this paper proposes a methodological framework for constructing migration flow vector spaces. Two vector construction methods are introduced: 1) constructing interaction force fields based on the difference between interactions in two directions divided by their sum; 2) representing vectors based on absolute flow quantities to establish the vector space. After demonstrating the irrotational of the field and satisfying the Gauss (divergence)'s theorem, a scalar potential field is inferred through integration, facilitating the description of implicit flow patterns within geographic interaction matrices.

This research exploration can be further extended to different aspects of vector space, such as different vector expressions, vector aggregation methods (at the origin or destination), vector weighting approaches, and vector field superposition methods. By constructing different vector representations of OD flows and comparing their visual effects and inherent properties, we can explore their suitability for various research topics and applications. Furthermore, the migration space constructed in this study mainly focuses on inter-city or inter-state scales. There is potential to investigate vector fields and scalar potentials of intra-city commuting flows. This would provide valuable insights into urban spatial organization, city centers, urban boundaries, infrastructure planning, and public services, among other aspects.

References

- 1 Michael Batty. *The new science of cities*. MIT press, 2013.
- 2 Zhaoya Gong and Jean-Claude Thill. Vector field based approach to reconstruct and represent functional spaces with areal cartogram, 2017.
- 3 Xi Liu, Clio Andris, and Bruce A Desmarais. Migration and political polarization in the us: An analysis of the county-level migration network. *PloS one*, 14(11):e0225405, 2019.
- 4 Mattia Mazzoli, Alex Molas, Aleix Bassolas, Maxime Lenormand, Pere Colet, and José J Ramasco. Field theory for recurrent mobility. *Nature communications*, 10(1):3895, 2019.
- 5 Peter Nijkamp and Waldemar Ratajczak. Gravitational analysis in regional science and spatial economics: A vector gradient approach to trade. *International Regional Science Review*, 44(3-4):400–431, 2021.
- 6 David A Plane. Migration space: Doubly constrained gravity model mapping of relative interstate separation. *Annals of the Association of American Geographers*, 74(2):244–256, 1984.
- 7 Waldo Tobler. *Spatial interaction patterns*, 1975.
- 8 Jingyuan Wang, Jiahao Ji, Zhe Jiang, and Leilei Sun. Traffic flow prediction based on spatiotemporal potential energy fields. *IEEE Transactions on Knowledge and Data Engineering*, 2022.
- 9 Hu Yang, Minglun Li, Bao Guo, Fan Zhang, and Pu Wang. A vector field approach for identifying anomalous human mobility. *IET Intelligent Transport Systems*, 17(4):649–666, 2023.

Causal Effects Under Spatial Confounding and Interference

Jing Zhang ✉

School of Geographical Sciences, University of Bristol, UK

Abstract

Spatial causal inference is an emerging field of research with wide ranging areas of applications. As a key methodological challenge, spatial confounding and spatial interference can compromise the performance of standard statistical inference methods. In the current literature, there is a lack of appreciation of the connections between spatial confounding and interference. This could potentially lead to overspecialized silos of research. Therefore, we need further research to bridge such gaps theoretically, and to find creative solutions for complex spatial causal inference problems. This short paper offers a brief demonstration: It discusses the connections between spatial confounding and interference. An illustrative simulation study shows how commonly used approaches compare across four test scenarios. The simulation study is discussed with an emphasis on the promising performance of counterfactual prediction based inference methods.

2012 ACM Subject Classification Applied computing → Law, social and behavioral sciences

Keywords and phrases Spatial causal inference, confounding, interference, counterfactual

Digital Object Identifier 10.4230/LIPICs.GIScience.2023.91

Category Short Paper

Funding *Jing Zhang*: Funded by the UK ESRC Southwest Doctoral Training Partnership.

1 Introduction

Knowledge of cause and effect plays an important role in explaining past events and planning for future ones. Causal inference is, broadly speaking, the empirical quest for such knowledge. The last seventy years have witnessed the formation of statistical inference frameworks that revolutionised empirical approaches to causal inquiries. Most notably, we have the Potential Outcomes (PO) framework [9] which approaches the inference of causal effect via an analogy to randomised experiments. It would also be fitting to describe this progress as part of a wider intellectual movement propelled by mutually reinforcing forces such as the vogue of evidence-based policy, the availability of data, and the maturity of causal theories.

Spatial causal inference is causal inference in the presence of substantive spatial causal mechanisms. Here, space can be interpreted as either geographical or relational. Over recent years, spatial causal inference has emerged as an independent area of research. On the one hand this is motivated by empirical topics that are irreducibly spatial. For example, policing and neighborhood crimes, vaccination and disease spread, air pollution and health... This makes spatial causal inference a valuable methodological endeavour with real world impact. On the other hand, this is also characterised by unique analytical challenges associated with spatial causal mechanisms that cannot be simply conceptualised as standard randomised experiments.

This short paper aims to offer a synthesis of two key concepts in spatial causal inference with illustrative examples. The paper was motivated by my observation that, in the current literature, there is a lack of appreciation of the connections between key analytical concepts, which could potentially lead to overspecialised silos of methodological research. Despite recent efforts to document progress in spatial causal inference (e.g. [8]), we have more of an



© Jing Zhang;
licensed under Creative Commons License CC-BY 4.0

12th International Conference on Geographic Information Science (GIScience 2023).

Editors: Roger Beecham, Jed A. Long, Dianna Smith, Qunshan Zhao, and Sarah Wise; Article No. 91; pp. 91:1–91:6

Leibniz International Proceedings in Informatics



LIPIC Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

assemble of techniques rather than a cohesive picture of the field. I believe the field will benefit from a consolidation of existing understandings of spatial causal problems as well as approaches to meeting the analytical challenges. In this short paper specifically, the focus will be placed on spatial confounding and interference. In the rest of the paper, I will first reflect on the two concepts. Then, with a simulation study, I will compare commonly used approaches across settings of spatial confounding and/or interference. The simulation will be discussed with an emphasis on the promising performance of counterfactual prediction based causal inference methods as an example of creative approaches that are able to engage multiple methodological topics.

2 Challenges in spatial causal inference

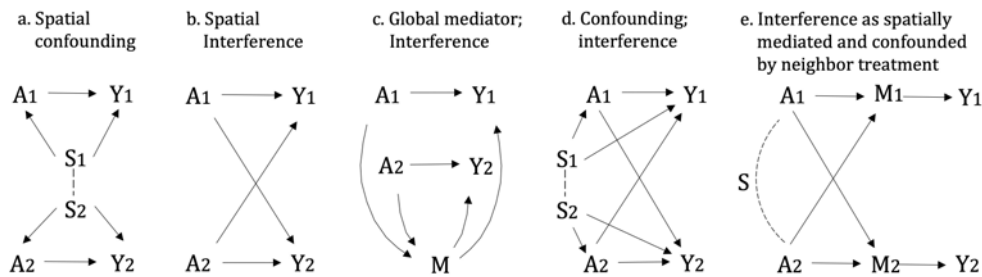
Spatial causal inference is characterised by its unique methodological challenges. Individual units are embedded in spatial contexts, and they interact in a spatially structured way. This tends to create more complex dependence structures than standard non-spatial causal inference methods admit. The resulted statistical problems are commonly captioned as spatial confounding, spatial interference, and spatial heterogeneity. Here, let's focus on spatial confounding and spatial interference. Specifically, I want to draw attention to the connections between spatial confounding and interference. Besides conceptual connections, the two problems often coexist in real world scenarios. Therefore, although methodological developments typically target one or the other, it is important that we understand how spatial causal inference methods engage with and perform under both spatial confounding and interference.

2.1 Spatial confounding

Confounding is a classic causal inference problem. Confounders influence both the treatment allocation and the outcome, and therefore not adjusting for the confounder admits a spurious correlation between the treatment and outcome variables. In spatial causal inference, we are particularly interested in confounders with significant spatial patterns (e.g. Figure 1.a), a condition which makes confounding adjustment amenable to spatial statistical techniques. The best way to think of spatial confounding is as a shorthand for spurious correlation due to omission of spatial variables. In recent literature, spatial confounding is mainly covered by the area of research on causal effects under unmeasured confounding. Under unmeasured confounding, the causal parameter in a PO model (typically the Average Treatment Effect, ATE) cannot be fully identified. Progress has been made on identification with propensity score matching (e.g. [2] [7]), using confounder proxy variables (e.g. [3]). For causal effect estimation, there are techniques to derive bounds for the causal parameter, for example, through sensitivity analysis (e.g. [1]), nonparametric bounding and interval estimates (e.g. [5]).

2.2 Spatial interference

In causal inference, 'interference' refers to the existence of dependence of an observational unit's outcome on the treatments of other units. In the PO framework, no interference is one of the basic assumptions, commonly known as one component of the Stable Unit Treatment Value Assumption. Spatial interference refers to scenarios of causal interference resulted from spatial interaction among the units. A typical case is treatment spillover, where a unit is exposed to a direct treatment as well as an indirect spillover treatment from its neighbours (e.g. Figure 1.b). This is what makes the interference problem unique, as we may



■ **Figure 1** Illustrative Directed Acyclic Graphs (DAG) for spatial confounding and interference. (Subscript denotes location.)

be interested in more than one causal estimands. The identifiability of causal effects under interference has been thoroughly investigated, among others, by Manski [6], and Forastiere [4]. Short of fieldwork-based exposure mapping to obtain true exposure levels, the estimation relies on strong restriction assumptions about the structure of causal interaction. Apart from interaction restrictions, the identification also relies on assumptions of no unmeasured confounding.

2.3 Common sources and shared solutions?

One way to appreciate how confounding and interference are connected is to reflect on the relationship between causal mechanisms and their reduced statistical representations. Although spatial confounding and spatial interference are conceptually distinct, they could be manifestations from the same underlying causal mechanism. In other words, it is possible that a given spatial causal mechanism, when translated as a statistical model, can present with either confounding or interference or both. As an illustrative case: When measuring the effect of vaccination on disease spreading, it can be conceptualised as an interference case (where the unvaccinated population receives a spillover protection from the vaccinated via mediation of group immunity, Figure 1.c); or it can be conceptualised as a confounding case (where the neighbourhood context of individuals confounds their actually received levels of protection as well as health outcome, Figure 1.a). Spatial confounding and interference can also coexist (e.g. Figure 1.d). With the example of neighbourhood crime rate interventions: The interference aspect is that intervention on one neighbourhood could affect crime rates of adjacent ones. There could coexist an element of confounding if intervention and crime rate variables are spatially distributed and a shared spatial trend creates a spurious dependence between them.

We can also try to understand the connection between spatial interference and confounding through the language of statistical causal inference. In a way, we can say that, a spatial interference problem is a spatial mediation problem wrapped within a confounding problem (e.g. Figure 1.e). After we peel away the confounding part with, for example, propensity score methods, the task of estimating direct and indirect effects is in spirit a task of estimating path specific causal effects. In the style of mediation analysis, the effect of direct treatment can be estimated conditional on indirect treatment levels and vice versa (e.g. [10]; [11]). In other words, an indirect effect is a causal effect mediated by the spatial interaction structure of the observational units, while the existence of such a structure usually also implies some degree of spatial confounding.

If spatial interference and confounding are so closely linked, what does this mean for methodology developments? To approach this question, first we have to better understand how the performance of existing methods generalises over spatial confounding and interference problems. So far, we have limited knowledge on this issue, as confounding and interference have been handled in separate strands of literature. To gain some insights, an illustrative simulation study is carried out.

3 Simulation study

The simulation study covers test scenarios of spatial confounding, spatial interference, and the coexistence of the two. Tested methods include two popular approaches to spatial causal inference: propensity score based adjustments, and spatial regression. Also tested is causal effect estimation based on counterfactual prediction of unobserved potential outcomes (also known as imputation based method). The counterfactual prediction approach is relatively new and has shown potential in addressing complex spatial causal inference problems.

3.1 Experiment design

The experiment is based on a basic setup. For the basic setup, the test dataset is generated in the following way: We have n observational units characterised by k covariates X^z drawn from a uniform distribution. Each unit inhabits a random location on a square. Its neighbours are defined as the set of units within a certain distance band. Its neighbourhood attributes X^g are represented by the average values of its neighbours' covariates. The assignment of direct treatment is independently determined by a unit's attributes X^z . The treatment Z is drawn from a Bernoulli distribution based on treatment propensity $e^z(X^z)$. A unit's outcome is determined only by its direct treatment status, $Z = 1$ treated and $Z = 0$ not treated. Accordingly, each unit has two potential outcomes, one of which is observed. The potential outcomes corresponding to direct treatment Z are $Y^z = Z * \tau^z + X^z * \beta + \epsilon$, $\epsilon^{i.i.d.} \sim N(0, 1)$, where τ^z is the average treatment effect parameter of interest. To reflect spatial causal inference problems, different spatial causal mechanisms are added to the basic setting. This includes the following test scenarios:

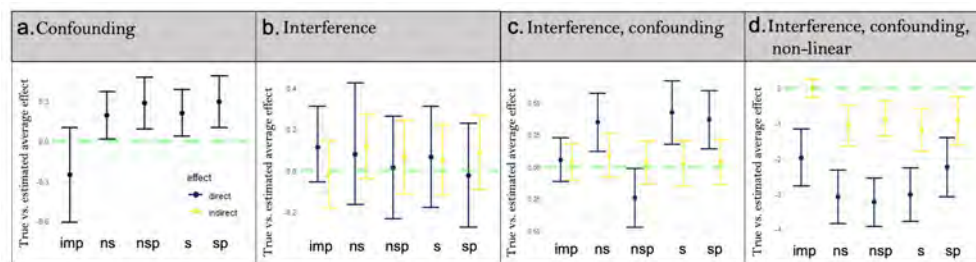
- (a) Spatial interference: In this scenario, besides direct treatment, a unit's outcome is also affected by its exposure to a neighbourhood treatment spillover G , $G = 1$ receiving spillover and $G = 0$ no spillover. The neighbourhood exposure G is determined by a unit's neighborhood covariate levels based on propensity $e^g(X^g)$. The marginal potential outcome corresponding to neighbourhood exposure G is $Y^g = \tau^g * G$, where τ^g is the average indirect treatment effect parameter. A unit's observed outcome is $Y = Y^z + Y^g$. Accordingly, each unit has four potential outcomes, one of which is observed.
- (b) Spatial confounding: To introduce spatial confounding, the X^z covariates are spatially smoothed, which introduces a common spatial pattern in the treatment and outcome variables.
- (c) Interference and confounding: A test scenario where both interference and confounding from scenarios (a) and (b) are present.
- (d) Non-linearity: On top of scenario (c), a non-linear function is used to generate the outcome variable.

The following list of causal inference methods are tested. They are denoted as:

- IMP: Imputing unobserved potential outcomes with non-parametric models, followed by inverse probability weighting to estimate average causal effects.

- NS: A baseline non-spatial PO model.
- NSP: Non-spatial PO model with propensity score adjustment.
- S: A baseline spatial model. The model takes the form of a spatial regression, as spatial econometric models are common in the estimation of spillover effects. The model is formulated as a PO model with spatially lagged treatment and confounder variables.
- SP: Model S with propensity score adjustment.

Some further clarifications: In the simulation, all the methods are implemented in their basic form for a fair comparison. While misspecification of the interference structure and inaccuracy of propensity score estimation are important sources of bias, in this experiment the test is kept simple. Where needed, true propensity scores and true interference network is used. For each scenario, the tests are run with sample size 1000, covariate dimension 5.



■ **Figure 2** Main results of simulation experiments.

3.2 Test results

Test results are reported in Figure 2. The four subplots corresponds to the four test scenarios. For each test scenario, the estimated average treatment effects from the five models are benchmarked on ground truth. A few findings from the results:

- (1) Across all test scenarios, comparing the performance of models ‘NS’ with ‘S’, and models ‘NSP’ with ‘SP’, we can see that the incorporation of spatial regression adjustment does not necessarily help to improve estimation accuracy. More generally, in applied cases it is difficult to verify the specification of spatial regression models, which can be a significant problem for causal inference tasks.
- (2) Comparing the performance of models ‘NS’ with ‘NSP’, and models ‘S’ with ‘SP’, the incorporation of propensity scores helps to adjust the estimates in the correction direction for most test scenarios. This is including the scenario with spatial interference and no confounding (Figure 2.b).
- (3) Comparing Figure 2.c with other scenarios, we can see that the coexistence of interference and confounding is challenging, as most models perform worse under this scenario. Meanwhile, compared with other models, the estimation accuracy of the counterfactual prediction based method ‘IMP’ does not deteriorate significantly when spatial confounding and interference coexist, suggesting a robustness of this approach.
- (4) Across all test scenarios, comparing the performance of ‘IMP’ models with others, we can see that the counterfactual prediction based method performs as well as the other methods in recovering the true causal effect. While propensity score based adjustments and spatial regression techniques are mainstream and have enjoyed decades of refinement, the counterfactual prediction approach is relatively new to spatial causal inference. Recently, the approach has been employed by Davis et al. [2] in a spatial confounding setting, and by Forastiere et al. [4] in a network interference setting. I believe the

counterfactual prediction approach is a promising direction of further methodological research. It is flexible enough to accommodate complex cases of spatial causal inference. And, it provides alternative ways to derive uncertainty quantification for models and parameters.



4 Conclusions

Spatial causal inference is an emerging field of research with wide ranging areas of applications. It is one of the methodological frontiers in the ongoing causal modelling movement. Complementary to existing review papers, this short piece offers a synthesis of two important concepts in spatial causal inference: spatial confounding, and spatial interference. A key message here is that: In the current literature, there is a lack of appreciation of the connections between core analytical concepts. This could potentially lead to overspecialised silos of research. Respectively, I believe several directions of research could benefit the field: Theoretically, we need further efforts on consolidating existing understandings of spatial causal problems and approaches to meeting the analytical challenges. Methodologically, counterfactual prediction is a promising direction of research which could potentially lead to flexible methods for complex spatial causal inference cases.

References

- 1 Carlos Cinelli, Daniel Kumor, Bryant Chen, Judea Pearl, and Elias Bareinboim. Sensitivity analysis of linear structural causal models. In *International conference on machine learning*, pages 1252–1261. PMLR, 2019.
- 2 Melanie L Davis, Brian Neelon, Paul J Nietert, Kelly J Hunt, Lane F Burgette, Andrew B Lawson, and Leonard E Egede. Addressing geographic confounding through spatial propensity scores: a study of racial disparities in diabetes. *Statistical Methods in Medical Research*, 28(3):734–748, 2019.
- 3 W Dana Flanders, Matthew J Strickland, and Mitchel Klein. A new method for partial correction of residual confounding in time-series and other observational studies. *American journal of epidemiology*, 185(10):941–949, 2017.
- 4 Laura Forastiere, Edoardo M Airoidi, and Fabrizia Mealli. Identification and estimation of treatment and interference effects in observational studies on networks. *Journal of the American Statistical Association*, 116(534):901–918, 2021.
- 5 Nathan Kallus, Xiaojie Mao, and Angela Zhou. Interval estimation of individual-level causal effects under unobserved confounding. In *The 22nd international conference on artificial intelligence and statistics*, pages 2281–2290. PMLR, 2019.
- 6 Charles F Manski. Identification of treatment response with social interactions. *The Econometrics Journal*, 16(1):S1–S23, 2013.
- 7 Georgia Papadogeorgou, Fabrizia Mealli, and Corwin M Zigler. Causal inference with interfering units for cluster and population level treatment allocation programs. *Biometrics*, 75(3):778–787, 2019.
- 8 Brian J Reich, Shu Yang, Yawen Guan, Andrew B Giffin, Matthew J Miller, and Ana Rappold. A review of spatial causal inference methods for environmental and epidemiological applications. *International Statistical Review*, 89(3):605–634, 2021.
- 9 Donald B Rubin. Formal mode of statistical inference for causal effects. *Journal of statistical planning and inference*, 25(3):279–292, 1990.
- 10 Eric J Tchetgen Tchetgen and Tyler J VanderWeele. On causal inference in the presence of interference. *Statistical methods in medical research*, 21(1):55–75, 2012.
- 11 Tyler J VanderWeele, Eric J Tchetgen Tchetgen, and M Elizabeth Halloran. Interference and sensitivity analysis. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 29(4):687, 2014.

Unlocking the Power of Mobile Phone Application Data to Accelerate Transport Decarbonisation

Xianghui Zhang  

SpaceTimeLab, Department of Civil, Environmental and Geomatic Engineering,
University College London, UK

Tao Cheng¹  

SpaceTimeLab, Department of Civil, Environmental and Geomatic Engineering,
University College London, UK

Abstract

Decarbonising transport is crucial in addressing climate change and achieving the Net Zero target. However, limitations arising from traditional data sources and methods obstruct the provision of individual travel information with comprehensive travel modes, high spatiotemporal granularity and updating frequency for achieving transport decarbonisation. Mobile phone application data, an essentially new form of data, can provide valuable travel information after effective mining and assist in progress monitoring, policy evaluation, and system optimisation in transport decarbonisation. This paper proposes a standardised methodology to unlock the power of mobile phone application data for supporting transport decarbonisation. Three typical cases are employed to demonstrate the capabilities of the generated individual multimodal dataset, including monitoring Londoners' 20-minute active travel target, transport GHGs emissions and their contributors, and evaluating small-scale transport interventions. The paper also discusses the limitations of mobile phone application data, such as issues surrounding data privacy and regulation.

2012 ACM Subject Classification Information systems → Geographic information systems; Applied computing → Transportation

Keywords and phrases Transport decarbonisation, Mobile phone application data, Application, London

Digital Object Identifier 10.4230/LIPICs.GIScience.2023.92

Category Short Paper

1 Introduction

Greenhouse Gases (GHGs) emissions, which contribute significantly to anthropogenic climate change, have emerged as a pressing global concern. The transport sector, a major source of GHG emissions, accounted for approximately 37% of worldwide CO_2 emissions from fuel combustion in 2021 [3]. Decarbonising transport is a multifaceted challenge involving the addressing of a variety of interrelated issues in policy, technology, and behavioural interventions. This includes encouraging model shifting, decreasing the high dependency on fossil fuels, expanding clean infrastructure investment, delivering transport interventions and regulations, and developing transport carbon credit trading market, etc [2]. In the meantime, numerous data-related challenges associated with these measures must also be resolved to support progress monitoring, intervention and policy evaluation, and the creation of effective policies and system optimisation for accelerating transport decarbonisation.

¹ Corresponding author



To be more specific, the availability of high-quality, detailed spatiotemporal data on travel behaviour, GHG emissions, and energy consumption is often limited, which hampers the creation of targeted interventions and the evaluation of their effectiveness. Secondly, tracking active travel data, which involves irregular trips and informal infrastructure, is challenging. This difficulty complicates the evaluation of modal shifts towards sustainable mobility and the optimisation of transport policies and infrastructure. Thirdly, inconsistencies in data collection methodologies and reporting standards across different jurisdictions prevent effective comparisons and the aggregation of data for regional or global analyses. These inconsistencies also influence the creation of decarbonisation targets and the assessment of progress in transport decarbonisation. Fourthly, near real-time data is vital for the operation of a transport carbon credit trading market and for promptly responding to (un)planned transport disruptions. While efforts have been made to address these data challenges, existing limitations continue to hinder progress towards transport decarbonisation.

Mobile phone data, a new form of data, holds several unique advantages, including high spatiotemporal granularity, large-scale coverage, passive data collection, real-time information, and integration with other geospatial data. Coupled with advancements in geospatial analysis and artificial intelligence, it becomes feasible to infer more holistic travel mode and personal activities information. For instance, its fine spatiotemporal granularity can inform infrastructure planning, from the expansion of cycling and public transit networks to the deployment of electric vehicle charging stations. Additionally, this data can bolster emissions models, aiding targeted mitigation efforts and enhancing our understanding of the connection between transport and emissions [4]. While mobile phone data offers substantial potential, its full utilisation to accelerate transport decarbonisation remains a significant challenge. The intricacies of processing massive data, the lack of standardised methodologies, and privacy concerns are areas that require further exploration and research.

This paper mainly introduces how do we develop a standardised methodology for unlocking the power of mobile phone application data and applying it to support transport decarbonisation. In the next section, we will introduce the superior characteristics of mobile phone application data and the used dataset. The third section will introduce the proposed methodology for mining the mobile phone application dataset and its potential applications. The fourth section will demonstrate the applications in monitoring active travel, estimating transport GHGs emissions, and evaluating transport interventions. In the last section, we will draw conclusions and discuss the potential limitations of mobile phone dataset.

2 Mobile phone application dataset

With the widespread adoption of smartphones and an increasing reliance on mobile networks, mobile phone data has become an abundant and valuable resource for researchers, businesses, and policymakers. Mobile phone data is derived from two sources, including cellular tower data and mobile phone application data. Cellular tower data is inferred from the connection of mobile devices to cell towers. Mobile phone application data is generated from built-in sensors and collected by popular public applications equipped with location-based services. In addition to the general characteristics, mobile phone application data has three advantages over cellular tower data, which enables a more detailed understanding of mobility patterns, travel modes, and location-based activities:

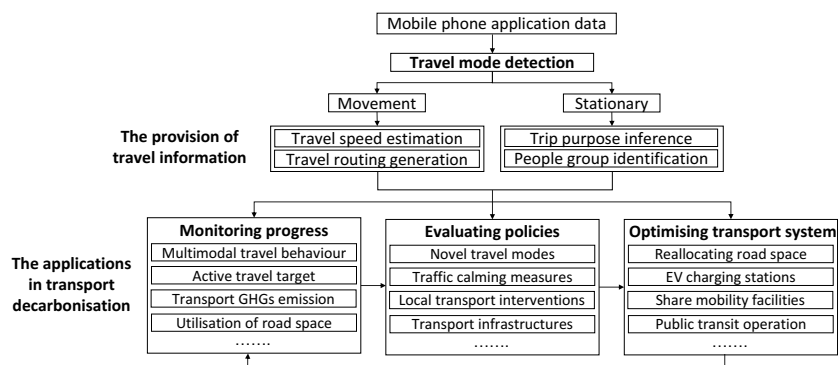
- Higher spatiotemporal granularity: mobile phone sensors (e.g., GPS) can provide more accurate location information compared to cellular network-based data, which relies on cell tower triangulation.

- Rich variety of sensors: mobile phone is equipped with numerous sensors, including accelerometers, gyroscopes, and magnetometers, which can provide additional context and information about individual activities, such as speed and heading.
- Higher sample rate: mobile phone application usually has higher collect data at specific time intervals or based on user-triggered events, allowing for the collection of high-frequency data.

In this study, we utilise a mobile phone application dataset from Location Science AI, which is collected and combined from more than 50 popular mobile phone applications. This dataset encompasses about 1 million+ unique devices in 2020 and 2021 in the UK. On average, we have about 80 data points per device per day, with the horizontal accuracy being approximately 21.7 meters. This effectively alleviates the potential sample bias in the mobile phone application data from a single source and increases the data sample per device. The obtained dataset, collected from the UK from January 2019 to the present (live feeding), provides a robust approach to investigating transport decarbonisation across multiple periods.

3 Methodology

This research presents a standardised methodology to harness the power of mobile phone application data for accelerating transport decarbonisation (**Figure 1**). The methodology comprises two main steps: (1) the provision of travel information. (2) the potential applications.



■ **Figure 1** the methodology for unlocking the power of mobile phone application dataset.

Given the raw mobile phone application dataset does not offer any travel-related information, it is essential to adopt and improve novel algorithms to mine travel information from it. In this framework, travel mode detection plays a pivotal role in providing travel information. We improve the moving window SVM model and combine spatial analysis for accommodating the massive data volume of mobile phone application dataset [1, 5]. Seven typical travel modes are detected, including car, bus, train, tube, cycle, walk, and stationary. After classifying the raw dataset into movement and stationary groups, we first further deliver individual travel information (e.g., speed and routing), which can be further aggregated to street-level traffic flow or regional-level OD flow. Besides, trip purpose (e.g., working and shopping) and people groups (i.e., residents, people with trip attractions, pass-through people) can be identified based on machine learning and spatial analysis methods by combining other spatial datasets (e.g., land use and POIs).

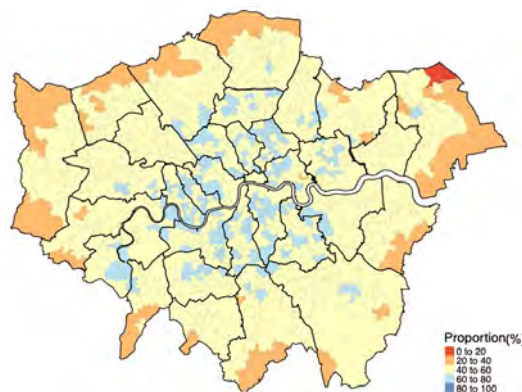
With detailed and comprehensive travel and activity information, many specific challenges in progress monitoring, policy evaluation, and transport system optimisation can be addressed. Mobile phone application dataset and this standardised methodology become the linkages between these three key aspects, thereby accelerating the achievement of transport decarbonisation. It should be noted that this framework only presents a part of travel-related information extraction and its applications, which could be further expanded to other unexplored fields.

4 Applications

In this study, we choose London as the study area and demonstrate three applications of transport decarbonisation based on mobile phone application data, specifically: active travel target achievement, transport GHGs emissions estimation, and transport intervention evaluation.

4.1 Monitoring the progress of active travel target

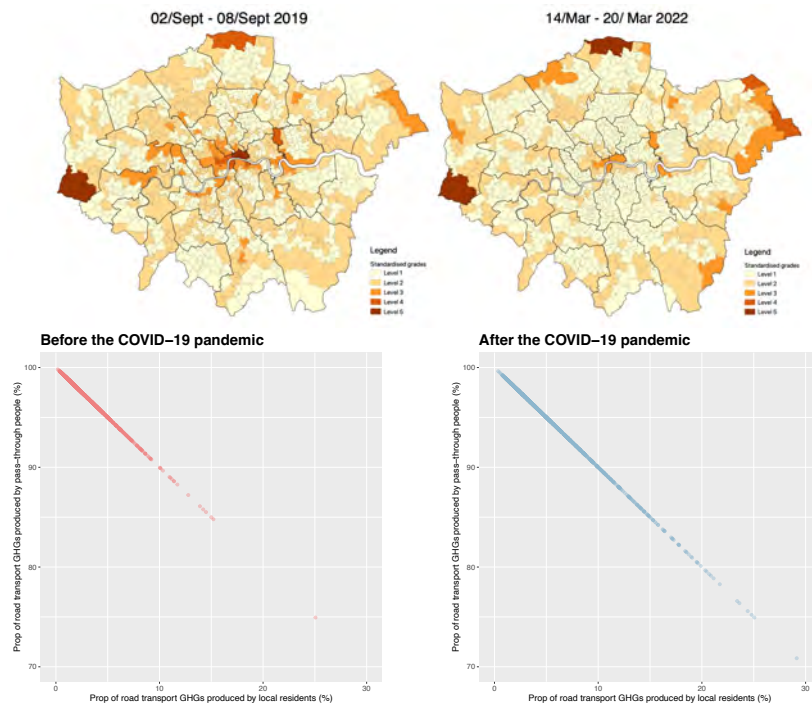
According to the Mayor of London's Transport Strategy, all Londoners should engage in at least 20 minutes of active travel by 2041. This means that local residents should be able to access essential services, amenities, and recreational opportunities within a 20-minute walk or cycle from their homes. This approach encourages people to use active and sustainable modes of transport, instead of relying on private cars. Consequently, it's important to monitor progress towards these targets to better direct sustainable mobility infrastructure development and related intervention formulation. However, continuous monitoring of active travel, including cycling and walking, is challenging with traditional statistics and surveys. Mobile phone application data could easily monitor local residents' active travel and assess the target achievement, making it simpler to optimise infrastructure investment and transport interventions. **Figure 2** presents the achievement of 20-minute active travel target at Middle Super Output Area (MSOA) level in London. Most MSOAs have not fully achieved the target, and MSOAs located in inner London have better progress in target achievement.



■ **Figure 2** the achievement of 20-minute active travel targets in London.

4.2 Estimating transport GHGs emissions

Estimating transport GHG emissions is a critical task for understanding the environmental impact of transport systems and for developing effective decarbonisation strategies. However, due to the complex nature of transport systems, the variability in vehicle technologies,



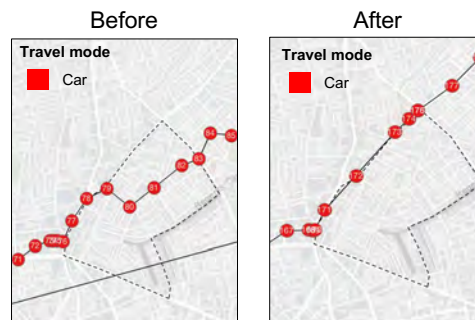
■ **Figure 3** The spatial distribution and contribution of transport GHGs emissions.

and user behaviours, estimating transport GHG emissions can be challenging. With the provided travel information, transport GHGs emissions can be further estimated by using distance-based estimation methods. Besides, the allocation of transport GHGs emissions could be further conducted based on travel modes and people groups derived from the mobile phone application dataset. **Figure 3** illustrates the transport GHGs emissions at MSOA level, as well as the shares of contributors (i.e., residents versus pass-through people) in London, both before and after the COVID-19 pandemic. The transport GHGs emission in inner London is no longer significantly higher than those in the outer, and the proportion of local transport GHG emissions produced by residents is higher than pre-pandemic. These changes may be attributed to shifts in travel behaviour and lifestyle, such as the increased prevalence of remote work. The findings will aid in formulating decarbonisation policies and foster the development of carbon credit trading markets.

4.3 Evaluating transport interventions

Transport interventions are crucial in promoting transport decarbonisation and steering towards more sustainability. They aim to reduce GHG emissions by endorsing low-carbon transportation modes, improving energy efficiency, and encouraging behavioural change. Key interventions include enhancing public transportation, supporting active travel modes, incentivising electric vehicle adoption, and implementing policies such as congestion pricing and low-emission zones. After the outbreak of the COVID-19 pandemic, London rapidly introduced Low Traffic Neighbourhood Scheme (LTN) across many boroughs, with the goal of encouraging active travel, preventing traffic through traffic, and keeping social distances. Given the scheme's small scale (approximately 1 km²) and uncertain implementation period, traditional datasets struggle to capture its impacts. Fortunately, the high spatiotemporal

granularity of mobile phone application dataset enables us to unpack LTNs' impacts. **Figure 4** illustrates the re-routing of individual driving routes after introducing St Peter LTNs in London. According to the travel information delivered from mobile phone data, it is easy to track the redistribution of multimodal traffic flow and road space usage. This is a powerful approach for authorities to evaluate transport interventions and identify potential side-effects.



■ **Figure 4** the re-routing of an individual driving route.

5 Conclusions

This study demonstrates how to unlock the power of mobile phone application data to accelerate transport decarbonisation and its potential applications. We reviewed the data challenges in transport decarbonisation and proposed a methodology to overcome them. We present three applications demonstrating how mobile phone application data can facilitate progress monitoring, transport GHGs emissions estimation, and policy and intervention evaluation. While mobile phone application data holds significant potential to support transport decarbonisation, two key limitations must be addressed in the future.

Firstly, while mobile phone application data offers valuable insights, it also contains sensitive information about individuals' movements and behaviours. It's crucial to strike a balance between utilising the potential of this data and ensuring privacy.

Secondly, data protection regulations, like the General Data Protection Regulation (GDPR) in the European Union, have been proposed or implemented worldwide to address growing data privacy concerns. These regulations dictate strict compliance regarding data collection, storage, and processing, which in turn pose new challenges to data availability and application. To mitigate these challenges, advanced data processing and management strategies, such as federated learning, should be developed or applied.

References

- 1 Adel Bolbol, Tao Cheng, Ioannis Tsapakis, and James Haworth. Inferring hybrid transportation modes from sparse GPS data using a moving window SVM classification. *Computers, Environment and Urban Systems*, 36(6):526–537, 2012. doi:10.1016/j.compenvurbsys.2012.06.001.
- 2 Department for Transport. Decarbonising Transport - A Better, Greener Britain. Technical report, United Kingdom Government, 2021.
- 3 IEA. Transport, IEA. Technical report, International Energy Agency, 2022.
- 4 Peilin Li, Pengjun Zhao, and Christian Brand. Future energy use and CO2 emissions of urban passenger transport in China: A travel behavior and urban form based approach. *Applied Energy*, 211(October 2017):820–842, 2018. doi:10.1016/j.apenergy.2017.11.022.
- 5 Xianghui Zhang and Tao Cheng. The impacts of the COVID-19 pandemic on multimodal human mobility in London: A perspective of decarbonizing transport. *Geo-spatial Information Science*, 00(00):1–13, 2022. doi:10.1080/10095020.2022.2122876.

The Ethics of AI-Generated Maps: DALL · E 2 and AI's Implications for Cartography

Qianheng Zhang ✉

HGIS Lab, Department of Geography, University of Washington, Seattle, WA, USA

Yuhao Kang¹ ✉ 

GeoDS Lab, Department of Geography, University of Wisconsin-Madison, WI, USA

Robert Roth ✉ 

Cartography Lab, Department of Geography, University of Wisconsin-Madison, WI, USA

Abstract

The rapid advancement of artificial intelligence (AI) such as the emergence of large language models ChatGPT and DALL · E 2 has brought both opportunities for improving productivity and raised ethical concerns. This paper investigates the ethics of using artificial intelligence (AI) in cartography, with a particular focus on the generation of maps using DALL · E 2. To accomplish this, we first created an open-sourced dataset that includes synthetic (AI-generated) and real-world (human-designed) maps at multiple scales with a variety of settings. We subsequently examined four potential ethical concerns that may arise from the characteristics of DALL · E 2 generated maps, namely inaccuracies, misleading information, unanticipated features, and irreproducibility. We then developed a deep learning-based model to identify those AI-generated maps. Our research emphasizes the importance of ethical considerations in the development and use of AI techniques in cartography, contributing to the growing body of work on trustworthy maps. We aim to raise public awareness of the potential risks associated with AI-generated maps and support the development of ethical guidelines for their future use.

2012 ACM Subject Classification Human-centered computing → Human computer interaction (HCI)

Keywords and phrases Ethics, GeoAI, DALL-E, Cartography

Digital Object Identifier 10.4230/LIPIcs.GIScience.2023.93

Category Short Paper

Related Version *Full Version*: <https://arxiv.org/abs/2304.10743> [11]

Supplementary Material *Dataset*: <https://github.com/GISense/DALL-E2-Cartography-Ethics>
archived at `swh:1:dir:a9d23d429831d2625a02551e0b9b1bb7131a0431`

Acknowledgements The authors would like to express their sincere gratitude for the support received from Dr. Song Gao at the GeoDS Lab, University of Wisconsin-Madison, Dr. Bo Zhao, and Yifan Sun at the HGIS Lab, University of Washington.

1 Introduction

Cartographers long have recognized the significance of developing ethical and trustworthy maps, i.e., maps that truthfully depict geographic information while minimizing the introduction of misinformation or bias [15, 6]. With the rapid advancements in Artificial Intelligence (AI), the use of AI in map-making has brought both opportunities and concerns [12, 9]. On the one hand, (Geo)AI techniques can facilitate map creation processes and even have

¹ Corresponding author: yuhao.kang@wisc.edu



demonstrated the potential to support human creativity in cartographic design. For instance, cartographers have employed (Geo)AI to support cartographic design decisions on the artistic aspects of maps such as map style transfer [10, 3], map generalization [4, 18], and map design critique [2]. On the other hand, despite its promise, cartographers have expressed ethical concerns about the uncertainty and opacity (i.e., machine learning and deep learning models often are considered as “black-boxes”) of AI for generating maps [21, 9]. As [6] asks: “How much should we trust a machine-generated map?”

Recently, generative models such as ChatGPT and DALL · E 2 have attracted significant public attention [16, 19]. These generative language models have demonstrated impressive capabilities in tasks such as language generation and image synthesis. Yet, they have fueled debates surrounding the ethical concerns related to the development of generative AI [13, 22]. The advancement of these generative AI models also raises critical questions about the future of labor [20] and the consequences of unbridled technological development [19, 14]. Therefore, there is an urgent need for careful consideration and ethical evaluation of AI technologies in myriad domains to ensure their responsible and beneficial use.

As cartographers and geographers, we have a particular interest in investigating the ethical implications of maps created by these advanced generative models. The emergence of powerful tools such as DALL · E 2 has made it increasingly accessible to generate high-quality map images by providing specific prompts. However, this also has introduced new challenges related to the accuracy and trustworthiness of these synthetic maps generated by AI. While these maps may look realistic, they also may contain inaccuracies or be influenced by biases embedded in the AI models, resulting in the proliferation of potentially meaningless, and, at worse, harmful maps online [17]. To address these issues, it is necessary to build solutions for detecting and mitigating the risks associated with using such maps, as suggested by [21]. Hence, it is crucial to offer timely detection of “fake” maps to assess the trustworthiness of web maps and minimize the potential negative impacts associated with their use.

To this end, we aim to investigate the use of AI in generating maps and the associated ethical implications of AI-generated maps. We ask the following two fundamental questions: (1) What potential ethical concerns arise from the characteristics of maps generated by DALL · E 2? and (2) How can AI-generated maps be identified to ensure their trustworthiness on web maps? To accomplish this, we first created a dataset that contains synthetic maps generated by DALL · E 2 with diverse prompts at multiple spatial scales (hereafter referred to as *AI-generated maps*). We also collected real-world maps using search engines (hereafter referred to as *human-designed maps*). In addition, we trained a deep learning-based model capable of identifying AI-generated maps. In this paper, we hope to use this study to apply ChatGPT-like generative models (e.g., DALL · E 2) in cartography. Our research contributes to the growing body of work on trustworthy maps and the ethics of cartography by highlighting the importance of ethical concerns in the development and use of AI techniques with cartography for the public.

2 Data and Methodology

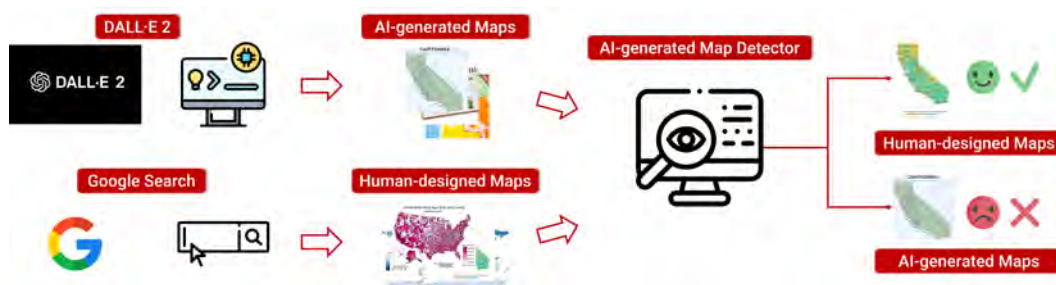
2.1 Construction of Dataset

We first created an AI-generated map dataset using the DALL · E 2 that relies on prompts to generate images. Specifically, we generated the maps using the following prompt format:

“A *{MapType}* of *{Region}* on *{Place}* with *{Description}*”

This format allows us to specify the type of map, the region of maps, the location where the map is put, and additional descriptive information for the AI model to generate a corresponding map.

The first two parameters for all prompts, *MapType* and *Region*, are required, while *Place* and *Description* are optional. For instance, to generate a United States choropleth map that is placed on the desk in warm colors, the prompt could be: “A choropleth map of United States with warm colors”. Then, we randomly selected options and combines them to generate a diverse set of maps covering various regions and themes. We have made the dataset openly available on GitHub at: <https://github.com/GISense/DALL-E2-Cartography-Ethics>. As a comparison, we developed a Python web scrapper to collect maps from the Google search engine at the same levels and administrative regions. To do so, we entered a search query in the format “*{Region}* maps”, such as “United States maps”. We adopted such a strategy used in prior studies to construct map datasets for country and continent levels [5, 8]. Regarding images at the state level, we directly utilized the dataset released from [8].



■ **Figure 1** The computational framework of this study.

2.2 Development of AI-generated Map Detector

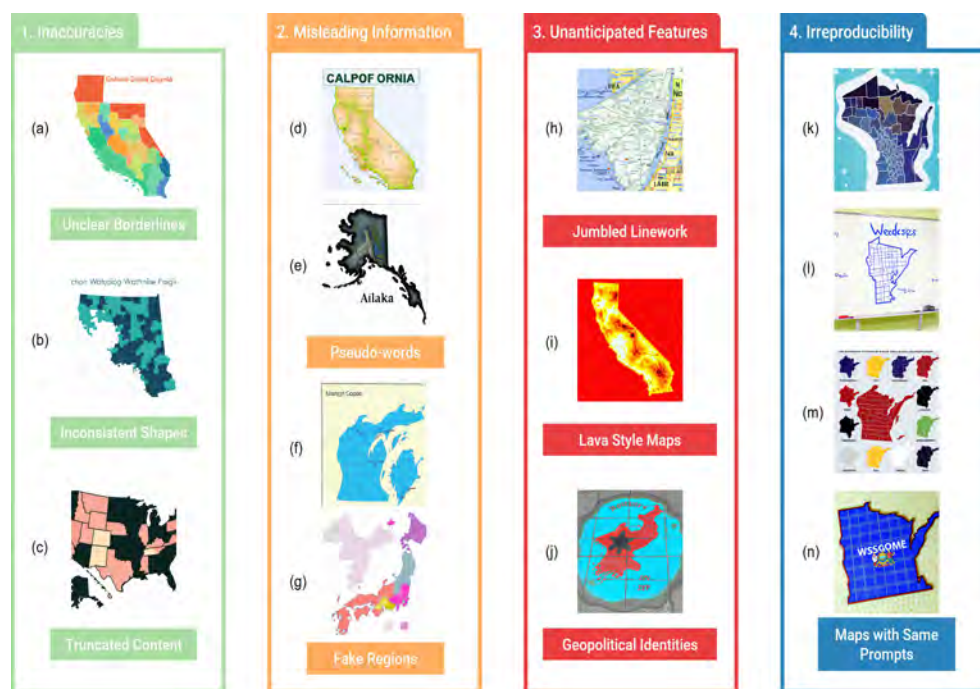
Based on the datasets, we developed an AI-generated map detector that can identify AI-generated maps which may offer potential solutions for creating trustworthy maps. Such a detector was developed based on a ResNet-18 model [7], a Deep Convolutional Neural Network (DCNN) that has been widely used in computer vision tasks due to its outstanding performance. Given its mature and high accuracy in image classification, we utilized the ResNet model in our study to classify maps as either generated by AI or created by humans. To train the model, we combine the two datasets, namely, AI-generated maps and human-designed maps, and input them into the ResNet-18 model.

3 Results

3.1 Ethical Issues of AI-generated maps

Based on our qualitative observations of AI-generated maps, we summarize four potential ethical concerns of such maps: inaccuracies, misleading information, unanticipated features, and the inability to reproduce results. We have included several examples of these map characteristics in figure 2, and we summarize our definitions of these characteristics below.

The inaccuracies observed in AI-generated maps are symbolized by unclear shapes of areas. Specifically, AI-generated maps may have unclear and distorted borderlines between different regions (e.g., states, and counties) or even unreasonable deformation of a certain



■ **Figure 2** Example AI-generated maps: (1) inaccuracies, (2) misleading information, (3) unanticipated features, and (4) irreproducibility.

area. At the same time, AI-generated maps from DALL-E 2 are limited to square output shapes since users cannot set the scale of images and they typically only display certain content of a region.

Moreover, AI-generated maps also can produce misleading information. AI-generated maps also can contain pseudo-words, non-existent provinces, or symbols, which create a false impression of the current map with a given input prompt. These features potentially can lead to the spread of misinformation or the distortion of popular notions of reality and have unintended geopolitical consequences and raise significant ethical concerns.

In addition to the inaccuracies and misleading information, AI-generated maps may create unexpected or unanticipated features. For instance, AI-generated maps are unaware of the underlying geographic processes that lead to repeated patterns in the landscape, particularly for the build environment in our study, resulting in distorted polygons or depicting a heat map as lava. The presence of polygons or lava suggests that the model may have misunderstood the meaning of the prompt. In addition, AI models may generate specific themes of maps that reflect certain geopolitical identities, even if not input in keyword prompts. Further work is needed to evaluate the degree to which AI-generated maps may stoke nationalism and thus reinforce xenophobic or otherwise biased geopolitical discourse.

Finally, AI-generated maps cannot be reproduced even with the same prompt. Due to the randomness inherent in the generation process of DALL · E 2, it is impossible to generate two maps that have the exact same map content, map shapes, map styles, or overall layouts. Without greater reproducibility, cartographic research on GeoAI cannot be validated or replicated, and therefore pose ethical questions about the effectiveness of a conventional, science-based peer-review system, and what “counts” as knowledge and scholarship more broadly. From a technical perspective, the model may reproduce the same outputs if we have the same hyperparameters (e.g., random seed, steps, prompts, weights). However, DALL · E 2 is not currently open-sourced and therefore is not reproducible at this time.

3.2 System Results of our AI-generated Map Detector

We evaluated the performance of our deep learning-based AI-generated map detection model on the test set. Based on the results, we computed four commonly used metrics in machine learning, namely, accuracy, precision, recall, and F1 score, to measure the performance. The system achieved an accuracy of 0.908, precision of 0.87, recall of 0.878, as well as an F1 score of 0.874 on the testing dataset. These metrics suggest the system is robust and effective in distinguishing between human-generated and AI-designed maps.

4 Discussions and Conclusions

While generative AI such as DALL·E 2 and ChatGPT have the potential to assist the cartographic design process, they also raise significant ethical concerns. In this paper, we present an AI-generated map dataset using DALL·E 2 and investigate the potential ethical issues associated with AI-generated maps based on their characteristics. The findings reveal that despite their promises, such AI-generated maps may deliver inaccurate and misleading information, contain unanticipated features, and lack reproducibility. In addition, we develop an AI-generated map detector with deep learning that can identify whether a map is generated by humans or by AI. This map detector is intended to be used in various applications, such as identifying potential cases of AI-generated maps being used to spread misinformation on online social media platforms. Inaccurate or misleading maps, whether intentionally or unintentionally created, may cause significant negative impacts, particularly in sensitive political or cultural contexts. It is possible for this map detector to help prevent the spread of misinformation and reduce the potential harm caused by AI-generated maps.

We acknowledge several limitations that are worth examining in the future. First, the dataset we collected in this paper was limited in geographic coverage and diversity since more diverse characteristics are required for the generalizability of our findings. Second, this paper has only investigated the maps generated by DALL·E 2 while numerous other models have been available that can produce maps. In this early stage of the application of AI technology, it is also important to explore more options in prompts and other parameters. More settings and models allow for a wide range of cartographic applications in generative AI that can be incorporated into future studies.




The future of AI in cartography involves generating more accurate and visually appealing maps through the rapid evolution of (Geo)AI technology, an emerging field known as “GeoAI for cartography”, “CartoAI”, or “MapAI”. It’s critical for cartographers to collaborate with AI developers to ensure cartographer-in-the-loop developments by addressing the limitations and minimizing potential ethical concerns. Also, AI could be used to facilitate collaborative mapping efforts [1]. By integrating multiple data sources, AI may make maps more accessible. However, the potential ethical issues (e.g., bias, trustworthiness) should be monitored or reduced. Participatory mapping, incorporating local knowledge, could improve map accuracy, promote community engagement, and foster collaboration, while addressing ethical concerns.

References




- 1 Robert Chambers. Participatory mapping and geographic information systems: whose map? who is empowered and who disempowered? who gains and who loses? *The Electronic Journal of Information Systems in Developing Countries*, 25(1):1–11, 2006.
- 2 Taisheng Chen, Menglin Chen, A-Xing Zhu, and Weixing Jiang. A learning-based approach to automatically evaluate the quality of sequential color schemes for maps. *Cartography and Geographic Information Science*, 48(5):377–392, 2021.

- 3 Sidonie Christophe, Samuel Mermet, Morgan Laurent, and Guillaume Touya. Neural map style transfer exploration with gans. *International Journal of Cartography*, 8(1):18–36, 2022.
- 4 Azelle Courtial, Guillaume Touya, and Xiang Zhang. Deriving map images of generalised mountain roads with generative adversarial networks. *International Journal of Geographical Information Science*, 37(3):499–528, 2023.
- 5 Michael R Evans, Ahmad Mahmoody, Dragomir Yankov, Florin Teodorescu, Wei Wu, and Pavel Berkhin. Livemaps: Learning geo-intent from images of maps on a large scale. In *Proceedings of the 25th ACM SIGSPATIAL international conference on advances in geographic information systems*, pages 1–9, 2017.
- 6 Amy L Griffin. Trustworthy maps. *Journal of Spatial Information Science*, 2020(20):5–19, 2020.
- 7 Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- 8 Yingjie Hu, Zhipeng Gui, Jimin Wang, and Muxian Li. Enriching the metadata of map images: a deep learning approach with gis-based data augmentation. *International Journal of Geographical Information Science*, 36(4):799–821, 2022.
- 9 Yuhao Kang, Song Gao, and Robert Roth. A review and synthesis of recent geoai research for cartography: Methods, applications, and ethics. In *AutoCarto 2022*, 2022.
- 10 Yuhao Kang, Song Gao, and Robert E Roth. Transferring multiscale map styles using generative adversarial networks. *International Journal of Cartography*, 5(2-3):115–141, 2019.
- 11 Yuhao Kang, Qianheng Zhang, and Robert Roth. The ethics of ai-generated maps: A study of dalle 2 and implications for cartography. *arXiv preprint arXiv:2304.10743*, 2023.
- 12 Gengchen Mai, Weiming Huang, Jin Sun, Suhang Song, Deepak Mishra, Ninghao Liu, Song Gao, Tianming Liu, Gao Cong, Yingjie Hu, Chris Cundy, Ziyuan Li, Rui Zhu, and Ni Lao. On the opportunities and challenges of foundation models for geospatial artificial intelligence, 2023. [arXiv:2304.06798](https://arxiv.org/abs/2304.06798).
- 13 Scott McLean, Gemma JM Read, Jason Thompson, Chris Baber, Neville A Stanton, and Paul M Salmon. The risks associated with artificial general intelligence: A systematic review. *Journal of Experimental & Theoretical Artificial Intelligence*, pages 1–15, 2021.
- 14 David Mhlanga. Open ai in education, the responsible and ethical use of chatgpt towards lifelong learning. *Education, the Responsible and Ethical Use of ChatGPT Towards Lifelong Learning (February 11, 2023)*, 2023.
- 15 Mark Monmonier. *How to lie with maps*. University of Chicago Press, 2018.
- 16 OpenAI. Gpt-4 technical report, 2023. [arXiv:2303.08774](https://arxiv.org/abs/2303.08774).
- 17 Anthony C Robinson, Pyry Kettunen, Luciene Delazari, and Arzu Çöltekin. New directions for the state of the art and science in cartography. *International Journal of Cartography*, pages 1–7, 2023.
- 18 Yilang Shen, Tinghua Ai, and Rong Zhao. Raster-based method for building selection in the multi-scale representation of two-dimensional maps. *Geocarto International*, 37(22):6494–6518, 2022.
- 19 Eva AM van Dis, Johan Bollen, Willem Zuidema, Robert van Rooij, and Claudi L Bockting. Chatgpt: five priorities for research. *Nature*, 614(7947):224–226, 2023.
- 20 Ali Zarifhonarvar. Economics of chatgpt: A labor market view on the occupational impact of artificial intelligence. *Available at SSRN 4350925*, 2023.
- 21 Bo Zhao, Shaozeng Zhang, Chunxue Xu, Yifan Sun, and Chengbin Deng. Deep fake geography? when geospatial data encounter artificial intelligence. *Cartography and Geographic Information Science*, 48(4):338–352, 2021.
- 22 Terry Yue Zhuo, Yujin Huang, Chunyang Chen, and Zhenchang Xing. Exploring ai ethics of chatgpt: A diagnostic analysis. *arXiv preprint arXiv:2301.12867*, 2023.

Digital Injustice: A Case Study of Land Use Classification Using Multisource Data in Nairobi, Kenya

Wenlan Zhang   

Centre for Advanced Spatial Analysis, University College London, UK

Chen Zhong¹   

Centre for Advanced Spatial Analysis, University College London, UK

Faith Taylor   

Department of Geography, King's College London, UK

Centre for Advanced Spatial Analysis, University College London, UK

Abstract

The utilisation of big data has emerged as a critical instrument for land use classification and decision-making processes due to its high spatiotemporal accuracy and ability to diminish manual data collection. However, the reliability and feasibility of big data are still controversial, the most important of which is whether it can represent the whole population with justice. The present study incorporates multiple data sources to facilitate land use classification while proving the existence of data bias caused digital injustice. Using Nairobi, Kenya, as a case study and employing a random forest classifier as a benchmark, this research combines satellite imagery, night-time light images, building footprint, Twitter posts, and street view images. The findings of the land use classification also disclose the presence of data bias resulting from the inadequate coverage of social media and street view data, potentially contributing to injustice in big data-informed decision-making. Strategies to mitigate such digital injustice situations are briefly discussed here, and more in-depth exploration remains for future work.

2012 ACM Subject Classification Applied computing → Environmental sciences

Keywords and phrases Data bias, Digital injustice, Multi-source sensor data, Land use classification, Random forest classifier

Digital Object Identifier 10.4230/LIPIcs.GIScience.2023.94

Category Short Paper

Funding *Chen Zhong*: The research has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 949670).

1 Introduction

Land use classification is an essential part of resource distribution such as conducting infrastructure upgrading projects and services provision activities. It is widely accepted to classify land use types using remote sensing data with census, survey, or interview [3]. Despite providing high-accuracy information, the traditional classification methods have common disadvantages of being labour-intensive, time-consuming, low spatial resolution and requiring substantial financial resources, which create barriers for the Global South countries to apply [5]. It is crucial to offer cost-effective and easily accessible methods for land use classification to decision-makers in the Global South. This would enable underprivileged countries to receive timely and precise information required for emergency assistance provision.

¹ Corresponding author



© Wenlan Zhang, Chen Zhong, and Faith Taylor;
licensed under Creative Commons License CC-BY 4.0

12th International Conference on Geographic Information Science (GIScience 2023).

Editors: Roger Beecham, Jed A. Long, Dianna Smith, Qunshan Zhao, and Sarah Wise; Article No. 94; pp. 94:1–94:6

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Unlike traditional survey-based data, big data - referring to sensor-collected automatic data - has gradually become a low-cost, timely, cost-efficient supplement to the traditional data sources in the Global South countries [6]. As a by-product of advancing technology and digitalisation, new data sources (e.g., social media, street view image) are generally collected by Internet of Things sensors and smart devices in the form of social media data, street view data, and remote sensing data [9]. The datasets can contain various information such as geo-referenced text, images, and GPS signals. This information can be used to analyse people's social activity patterns, and even hence infer the land use types.

However, it is estimated that 37% of the global population remains to have restricted or no access to the internet, and the disconnected proportion is unsurprisingly high in Global South countries. Those with no access to smart devices or the internet are called 'digitally invisible' since they have less opportunity to generate data that could influence policy or benefit from data-informed analysis [2]. This data-caused discrimination, together with visibility and engagement with technology, was concluded as a data justice challenge by Prof. Linnet Taylor [8]. Data bias and the impact of digital injustice have created an obstacle to the application of big data. However, limited research has been conducted to verify digital injustice and to propose effective strategies for its mitigation. Therefore, this research aims to identify instances of digital injustice by performing a land use classification using multi-source publicly available data, with a case study of Nairobi, Kenya. The question of who constitutes the digitally invisible groups and where they reside remains an unresolved issue for future work.

2 Study Materials and Methodology

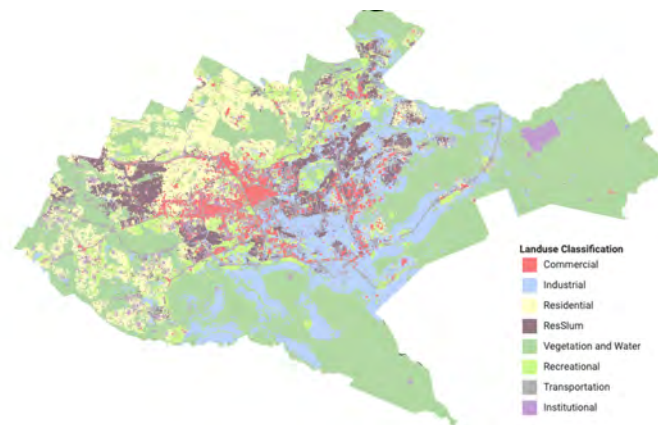
The case study city of Nairobi is the capital city of Kenya, which has been the economic centre of East Africa [4], experiencing overwhelming population growth and informal settlements expansion. These informal settlement areas accommodate more than 60% of the total population while occupying less than 5% of the city's residential land area [1]. Due to the rapid pace of development, the land use and extent of informal settlements can change significantly within a short period of time. Therefore, frequently updated land use data would be beneficial for local decision-makers.

The 2010 land use map shapefile of Nairobi, Kenya, created by Columbia University's Center for Sustainable Urban Development and obtained from the World Bank Data Catalog, served as the training dataset. However, due to the prolonged interval since its release and the swift pace of urban development in Kenya, various modifications were implemented based on field investigations and comparisons using Google Maps. The initial dataset encompassed 13 categories, which were subsequently condensed to 8 categories in accordance with the Nairobi land use policy, namely, commercial, industrial, residential, informal settlements, vegetation, water, recreational, transportation, and institutional.

Multiple sources of open sensor data were employed to conduct the research, and the relevant information is summarised in the table 1.

■ **Table 1** Data source and feature.

Data (abbr.)	Raster Data		Vector Data		
	Satellite images (R)	Night-time light (N)	Building footprints (O)	Social media posts (T)	Street view images (S)
Source	Sentinel-2 MSI	VIIRS-DNB	Google Open Building	Twitter posts	Mapillary
Information	Spatial resolution 10m	Spatial resolution 760m	Polygon	Text point	Image point
Feature selection	Bands, NDVI, NDWI, NDBI ²	Night Band	Building density	Tweet language and time	Object detected
Feature interpretation	Land physical char	Urban extent	Building char	Social activity	Sectional physical environment



■ **Figure 1** Land use map with 8 categories.

The raster data was resampled to a unified spatial resolution of 30 meters to provide detailed information suitable for community and city-level analysis. However, this resolution was chosen primarily for illustration purposes, and the accuracy trend is expected to perform similarly across different spatial units. Twitter posts were categorized into three categories: working/school, leisure time, and home time, based on whether the post was made on a weekday or weekend and the time of the post. The content of the posts was analysed using language detection techniques. A panoramic segmentation of the street view images was conducted using Detectron2, a pre-trained object detection algorithm developed by Meta. The processed vector dataset was then rasterized to a 30-meter resolution to align with the remote sensing data.

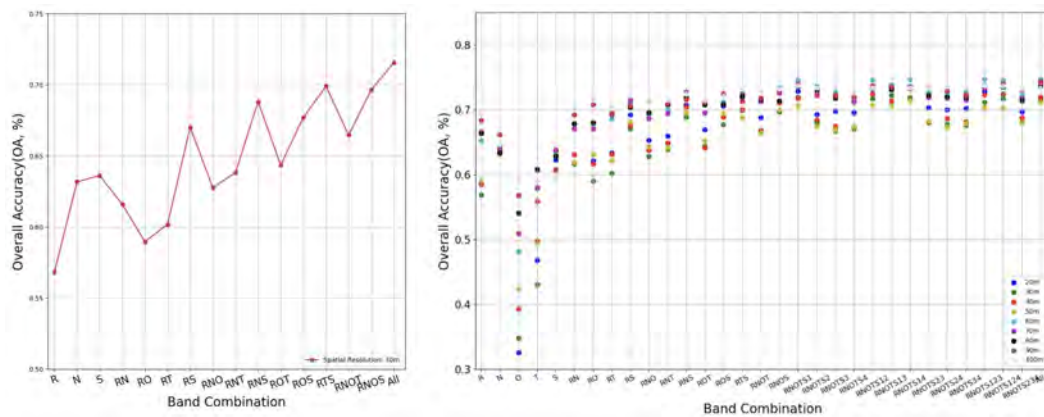
In this study, the random forest was employed as the benchmark classifier for land use classification for illustration purposes, since it has been widely applied and considered to be the most effective method for land use and land cover classifier [7]. Random forest is a supervised learning technique, whereby the classification categories can be allocated from the training dataset. A sample of 1000 pixels was randomly selected from each category and split into training and test sets. The forest number was set to 200. It is worth noting, however, that the selection of the classifier does not constitute the primary objective of this research, and other classifiers could be utilised in lieu of random forest. Although the overall accuracy of different data combinations may differ, significant modifications to the ranking of overall accuracy are not anticipated.

3 Result and Discussion

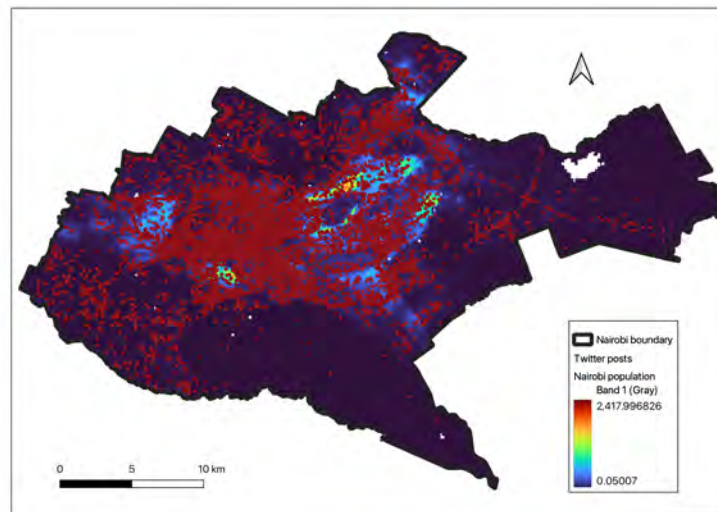
3.1 Land use map

The predicted Nairobi land use map with a 30m spatial resolution is presented in Figure 1. Figure 2(a) illustrates the change in OA with different data combinations. The combination of all datasets achieved the highest overall accuracy of 71.57%. As hypothesised, the aggregation of multiple data sources significantly enhanced the effectiveness of the OA of land use classification. This trend is consistent across different spatial resolutions, as shown in Figure 2(b).

² NDVI: Normalised Difference Vegetation Index, NDWI: Normalised Difference Water Index, NDBI: Normalised Difference Build up Index



■ **Figure 2** Land use classification (a) OA of 30m spatial resolution; (b) OA across spatial resolution.

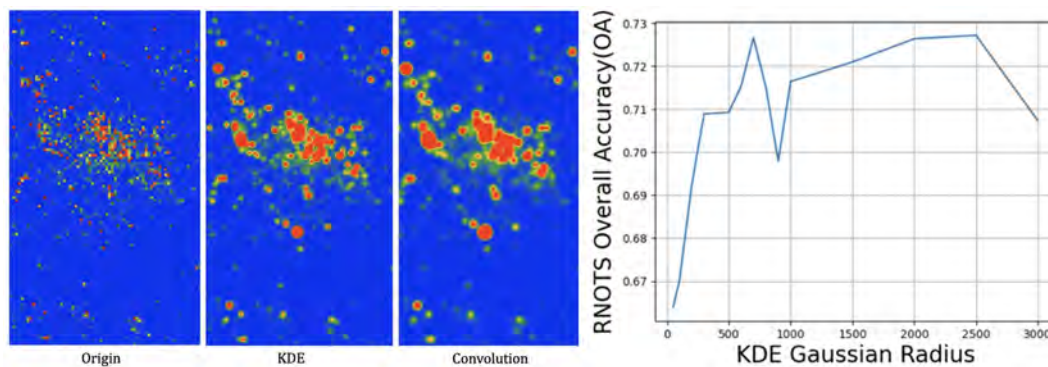


■ **Figure 3** Twitter data biased spatial distribution.

This performance could be attributed to the fact that data aggregation allowed for a full range of information to be revealed. The satellite images, night-time light images, and building density reflect the physical features of the land. In addition, social and economic features, such as different languages (English, Swahili, or others) found in Twitter posts, can provide insight into the people’s education levels and social connectivity, as they may use English exclusively for professional and outreach activities in commercial areas. Industrial areas are among the most commonly used places for Swahili, which could contribute to increased accuracy. Moreover, the presence of umbrellas in street view images could be used to directly infer the presence of a commercial area, as it is a unique indication of a local roadside market. These findings provide further evidence of the importance of using mobility data to identify social and economic features.

3.2 Mitigating data injustice

The results also highlighted the existence of low-data areas in Nairobi, as can be seen in Figure 3. Some highly populated area (colour in red and yellow in the background), especially the informal settlements of Kibera and Mathare, was not covered by Twitter post. After



■ **Figure 4** KDE and convolution (a) Regional performance (b) OA performance.

dividing the city into a grid with a spatial resolution of 30 meters, there were a total of 792,534 grid cells. However, only 307,632 cells had valid Twitter posts with identified language, which accounted for 38.82% of the total area. People who live in places where no data is collected are digitally invisible groups. The existence of digitally invisible groups would reduce classification accuracy, and potentially lead to biased decision-making.

The possible reasons for this uneven data distribution were: (1) genuinely less populated areas: the urban outskirts contain underdeveloped bare land and agricultural land. (2) low internet or smart device penetration: as mentioned before the rural area would have lower smartphone access. (3) a preference for other social media platforms: according to research done by Kepios (Kemp, 2022), the social media preference ranking: Facebook (42.6%), LinkedIn > (12.4%) > Instagram (10.7%) > Snapchat (7.5%) > Twitter (5.8%).

Tobler's First Law of Geography suggests that neighbouring areas are more similar than distant ones. Based on which, we assume that increasing the impact of a single data point could potentially cover nearby no-data areas and amplify the voices of digitally invisible groups. This could be implemented by performing a kernel density estimation (KDE), followed by a Gaussian convolution, as shown in Figure 4(a). The land use classification accuracy with all bands (at a spatial resolution of 30 m) increased from 57.68% to 70.24%. This result proved our assumption that nearby land use can be inferred using single data points.

Determining the impact range of a single data point remains a critical question. To understand the effect of distance on land use classification accuracy, an optimisation of the parameter has been plotted as shown in Figure 4(b). The optimal performance distance for Twitter posts in Nairobi was approximately 700 meters, resulting in an OA of 72.72%. This finding suggests that land use types tend to remain consistent within a 700-meter radius in Nairobi. However, it should be noted that this approach only addresses digital injustices within a specific range. As the distance increases, land use categories may differ significantly, and thus, data gaps for large data-missing areas cannot be inferred. Therefore, designing surveys and interviews as supplementary data collection to visualise the digitally invisible groups for large data-missing areas would be beneficial.

4 Limitation and Future Work

This project is subject to certain limitations that need to be acknowledged. Firstly, the findings highlight the presence of data bias and digital injustice, along with a brief analysis of their spatial extent. However, the quantitative spatial coverage and representativeness of


the sensor data were not fully explored, which leaves open questions about the demographic, spatial, and temporal distribution of the digitally invisible population. Consequently, only limited mitigation approaches were provided, and no information was provided about who should be the target group for the small data collection. The unresolved inquiries also include whether big and small datasets can represent different social groups and whether performing data fusion can be implemented to mitigate digital injustice. These questions will be further explored in the next phase of our research.

The predicted land use map may not fully capture areas with multiple functions due to the relatively coarse 30m spatial resolution. This limitation is caused by the limited computing capacity of GEE. However, for city-level decision-making, a 30m resolution is generally sufficient. To overcome this limitation for more granular analyses, one can zoom in to a smaller area or switch to another server.

References

- 1 Stefanos Georganos, Angela Abascal, Monika Kuffer, Jiong Wang, Maxwell Owusu, Eléonore Wolff, and Sabine Vanhuysse. Is it all the same? mapping and characterizing deprived urban areas using worldview-3 superspectral imagery. a case study in nairobi, kenya. *Remote Sensing*, 13, December 2021. doi:10.3390/rs13244986.
- 2 Justin Longo, Evan Kuras, Holly Smith, David M. Hondula, and Erik Johnston. Technology use, exposure to natural hazards, and being digitally invisible: Implications for policy analytics. *Policy and Internet*, 9:76–108, March 2017. doi:10.1002/POI3.144.
- 3 Darius Phiri, Matamayo Simwanda, Serajis Salekin, Vincent R. Ryirenda, Yuji Murayama, Manjula Ranagalage, Nadya Oktaviani, Hollanda A Kusuma, Tianxiang Zhang, Jinya Su, Cunjia Liu, Wen Hua Chen, Hui Liu, Guohai Liu, M. Cavour, H. S. Duzgun, S. Kemec, D. C. Demirkan, Radhia Chairat, Yassine Ben Salem, Mohamed Aoun, Zolo Kiala, Onesimo Mutanga, John Odindi, and Kabir Peerbhay. Sentinel-2 data for land cover / use mapping: A review. *Remote Sensing*, 12:12291, 2020.
- 4 Hang Ren, Wei Guo, Zhenke Zhang, Leonard Musyoka Kisovi, and Priyanko Das. Population density and spatial patterns of informal settlements in nairobi, kenya. *Sustainability 2020*, Vol. 12, Page 7717, 12:7717, September 2020. doi:10.3390/SU12187717.
- 5 Yan Shi, Zhixin Qi, Xiaoping Liu, Ning Niu, and Hui Zhang. Urban land use and land cover classification using multisource remote sensing images and social media data. *Remote Sensing*, 11:2719, November 2019. doi:10.3390/RS11222719.
- 6 Aiman Soliman, Kiumars Soltani, Junjun Yin, Anand Padmanabhan, and Shaowen Wang. Social sensing of urban land use based on analysis of twitter users' mobility patterns. *PLOS ONE*, 12:e0181657, July 2017. doi:10.1371/JOURNAL.PONE.0181657.
- 7 Swapan Talukdar, Pankaj Singha, Susanta Mahato, Shahfahad, Swades Pal, Yuei An Liou, and Atiqur Rahman. Land-use land-cover classification by machine learning classifiers for satellite observations—a review. *Remote Sensing 2020*, Vol. 12, Page 1135, 12:1135, April 2020. doi:10.3390/RS12071135.
- 8 Linnet Taylor. What is data justice? the case for connecting digital rights and freedoms globally. *Big Data and Society*, 4, December 2017. doi:10.1177/2053951717736335.
- 9 Linnet Taylor and Dennis Broeders. In the name of development: Power, profit and the datafication of the global south. *Geoforum*, 64:229–237, August 2015.

Exploring Map App Usage Behaviour Through Touchscreen Interactions

Donatella Zingaro ✉ 

Department of Geography, University of Zurich, Switzerland

Mona Bartling ✉ 

Department of Geography, University of Zurich, Switzerland

Tumasch Reichenbacher ✉ 

Department of Geography, University of Zurich, Switzerland

Abstract

Mobile map apps are rapidly changing the way we live by providing a broad range of services such as mapping, travel support, public transport, and trip-booking. Despite their widespread use, understanding how people use these apps in their everyday lives is still a challenge. In order to design context-aware mobile map apps, it is important to understand mobile map app usage behaviour. In this study, we employed a novel approach of recording touchscreen interactions (taps) on mobile map apps and combined them with users' distances from their homes to capture everyday map app usage. We analysed data from 30 participants recorded between February 2021 and March 2022 and applied two different data-driven analysis techniques to evaluate map apps usage. Our results reveal two distinct tapping signatures: a “home behaviour”, characterised by high interactions with map-related apps close to home, and a “travel behaviour”, defined by lower interactions scattered over a range of distances. Our findings have important implications for future work in this field and demonstrate the potential of our new approach for understanding mobile map app usage behaviour.

2012 ACM Subject Classification Human-centered computing

Keywords and phrases mobile maps, tappigraphy, cluster analysis, archetypal analysis, user-context, map-app usage

Digital Object Identifier 10.4230/LIPICs.GIScience.2023.95

Category Short Paper

Funding *Mona Bartling*: This research was funded by the Austrian Science Fund (FWF) J 4631-N.

1 Introduction

GIScience is facing complex challenges in the digital era, including the design of mobile map apps on smartphones. These apps provide various mainstream services such as navigation, travel assistance, public transport, and trip booking. However, it remains unclear how exactly people engage with these apps on a daily basis and how to effectively design such apps based on people's needs. Research on (mobile) map design primarily uses controlled experiments for the design evaluation, such as think-aloud methods in field or lab studies, limiting the ecological validity of the findings [6, 8]. Studies conducted in real-world settings are needed to understand mobile map app usage “in the wild.” To our best knowledge, to date, only one study matched this approach [7], by continuously collecting map-related usage data from users' mobile devices to identify map-usage scenarios and patterns (i.e. interaction patterns such as map-view manipulation, searching and finding a place on the map, navigating to a place). As the study was limited to a single map app (Google Maps), we argue that to understand the plurality of mobile map apps, we must explore more than just one app. Thus, new methodologies are needed to evaluate map usage behaviour in a real-world setting to understand a broader range of mobile map apps. This paper addresses



© Donatella Zingaro, Mona Bartling, and Tumasch Reichenbacher;
licensed under Creative Commons License CC-BY 4.0

12th International Conference on Geographic Information Science (GIScience 2023).

Editors: Roger Beecham, Jed A. Long, Dianna Smith, Qunshan Zhao, and Sarah Wise; Article No. 95; pp. 95:1–95:6

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

this need by introducing a novel approach, tappigraphy, which records usage behaviour with mobile apps by leveraging smartphone touchscreen interaction data (i.e. taps on a device display). Tappigraphy originated and is widely used in neuroscience to uncover behavioural patterns [1]. Recently, it has been shown to be applicable to GIScience as an ecological momentary assessment (EMA) tool to study mobile map usage behaviour [5]. Unlike other EMA methods, tappigraphy solely involves recording the taps on the screen of a smartphone and deliberately does not require the knowledge of any other private information about the individual (e.g. gender, nationality). Moreover, unlike lab-based investigations, recording touchscreen tapping data offers a new, unobtrusive way to observe human behaviour in everyday activities. Furthermore, it is not limited to selected apps and can be flexibly used to evaluate multiple apps in the aggregate. With this paper, we aim to employ tappigraphy in analysing map app usage behaviour in relation to the study participants' distance from their home location. By analysing the frequency of map-related taps at various locations to the participant's home, we intend to infer similar or different user behaviour pertaining to map apps.

2 Methodology

Through the MapOnTap app, we collected data for a minimum of two-weeks from thirty-eight participants. Data recording took place between February 2021 and March 2022. Participants were asked to install the free Android MapOnTap app on their smartphones. It is based on a tap counting app, which operates in the background on a smartphone. The recording of participants' phone sessions starts the moment they began unlocking the screen and continued until it was locked again. Within each phone session, tapping data on the active foreground apps were recorded as a series of timestamps, including the total number of taps, the start and end time, the apps used during each phone session, the participant's randomised ID code, the device ID (i.e. a generated code for each participant's device) and the Google Play Store app category associated with the used app. In addition to the tapping data, the app optionally records GPS coordinates. For the purpose of our study, we asked our participants to use their smartphones as usual and activate the MapOnTap app for at least two consecutive weeks. Participants were free to stop recording, turn GPS tracking off, or delete the app any time they wished. We did not collect any other information about the participants. The data collection was approved by the ethical board of University of Zurich.

From our initial dataset, we excluded three participants whose data collection period was less than two weeks. The duration of data collection for each individual, varied from a minimum of 14 days to a maximum of 313 days (M: 121 days, SD: 94 days). We applied different pre-processing steps to analyse our data. First, we calculated the Euclidean distances to the participants' home locations for each phone session. For the participants' home locations, we assumed that the most frequent coordinate pairs corresponded to the home of our participants. Given that part of the data collection occurred during the Covid-19 pandemic-related restrictions, it is reasonable to assume that the mobility patterns of our participants may have been influenced by pandemic-induced factors. In order to analyse the data, we first calculated the distances from participants' homes for each tap record, and then we aggregated the total number of taps for each app category. We used the category list from the Google Play Store as a reference for this process. We specifically selected the categories of "Maps and Navigation" (MN) and "Travel and Local" (TL), as they are the only two categories that are explicitly related to map apps. Subsequently, we excluded five participants from the study whose tapping data did not include any recordings for the MN

and TL app categories. Our dataset revealed 74,304 distance values ranging from 0 to 9,000 km from home. To identify and exclude any outliers, we applied the interquartile range method. As a result, we eliminated 2,121 extreme values from the dataset. The resulting distances ranged from 0 to 1,393 km from participants' home locations. Next, we calculated distance intervals by applying the Fisher-Jenks algorithm. With this, we were able to assign the recorded distances to 100 distance interval bins and label them with the median distance value of each bin. Our two final datasets consisted of 30 rows representing the final number of participants included in our analysis and 100 columns representing the number of taps corresponding to each distance bin that we computed for the two aforementioned app-related categories. Finally, the tap values were standardised by calculating the z-score.

3 Results and Discussion

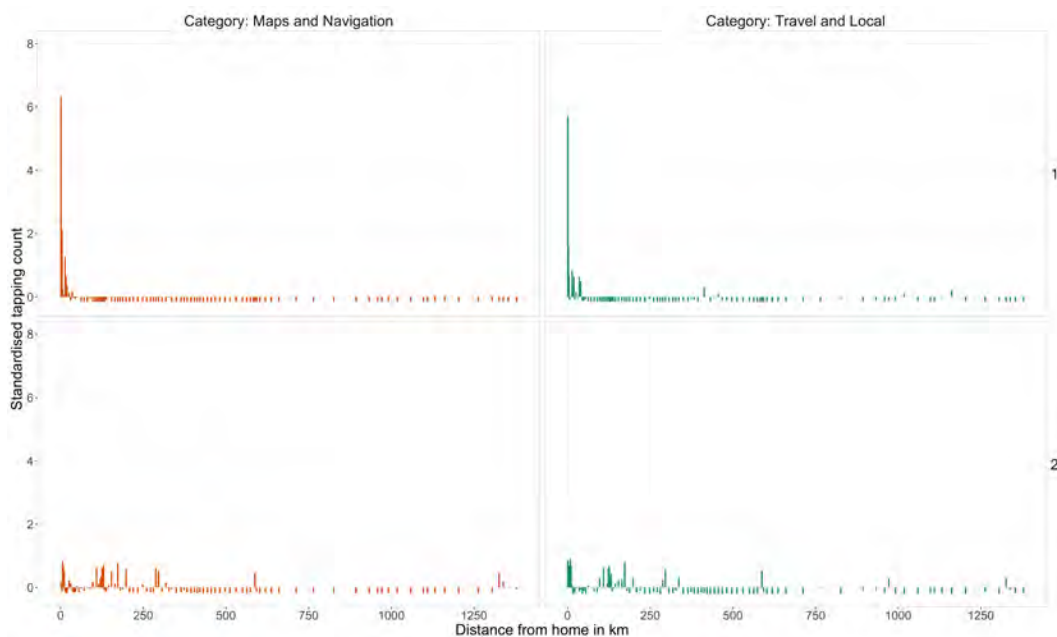
3.1 Descriptive Statistics on Tappigraphy Data

A total of 1087 unique apps were found in our dataset, catalogued in 33 categories according to Google Play Store (e.g. Social, Communication, etc.). Of these categories, only MN and TL refer to map-related apps were selected. We found 25 unique apps related to the MN category. For instance, navigation tools, mapping, and public transportation apps (e.g. Petal map, a mapping service from Huawei and apps of public railways companies, such as SBB). For the TL category, we identified 63 unique apps. For example, travel-booking tools, ride-sharing apps, trip management tools, and tour-booking apps (e.g. Booking, TripAdvisor, Publibike). Upon examining the total number of taps in our dataset, the TL category has, on average, more than twice as many taps recorded (M:3,798, SD:5,775) as the MN category (M:1,707, SD:4,218). Further, the TL category also had a greater maximum number of taps (132,923) than the MN category (59,734). The relatively high standard deviation can be attributed to the varying degrees of participation and data collection duration among participants, as the data collection period spanned almost a year. This unbalanced nature of the dataset is a trade-off of the study design, and may have impacted our results.

In terms of tap records in association with distances from home, the tapping data is not uniformly distributed but concentrated within a range of 200 km, with 84% of taps recorded for the MN category and 62% of taps for the TL category falling within this distance range.

3.2 Hierarchical Cluster Analysis (HCA) and Archetypal Analysis

Our main goal was to uncover potential map app usage patterns at varying distances from participants' home locations. To this end, we focused on two methods: HCA and Archetypal Analysis. HCA is an unsupervised algorithm that forms ordered subgroups, which can help individualise data clusters that are more closely or distantly related [4]. We employed Ward's criterion to optimise homogeneity within clusters by minimising the within-cluster sum of squares. HCA is typically represented by a dendrogram, where the height of branches represents the distance or dissimilarity between clusters. To determine the optimal number of clusters, we partitioned the dendrogram to maximise nodes' distances between the tree [4]. The HCA analysis resulted in two clusters, with cluster 1 comprising 17 participants, and cluster 2 comprising 13 participants. Figure 1 illustrates the mean standardised tapping counts of participants for both the MN and TL categories over distances from participants' home location of both clusters. It can be observed that participants in cluster 1 exhibit a strong interaction pattern within approximately 1 km of distance to their home location for both app categories. In contrast, cluster 2 is characterised by a more dispersed interaction

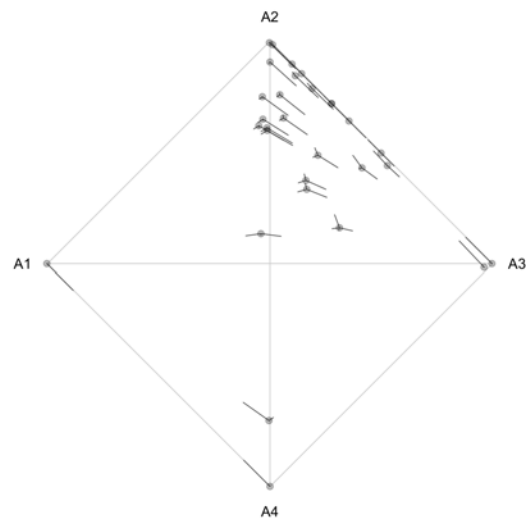


■ **Figure 1** Bar chart displaying the HCA clustered mean tapping counts of participants, for two categories: Maps and Navigation (left) and Travel and Local (right).

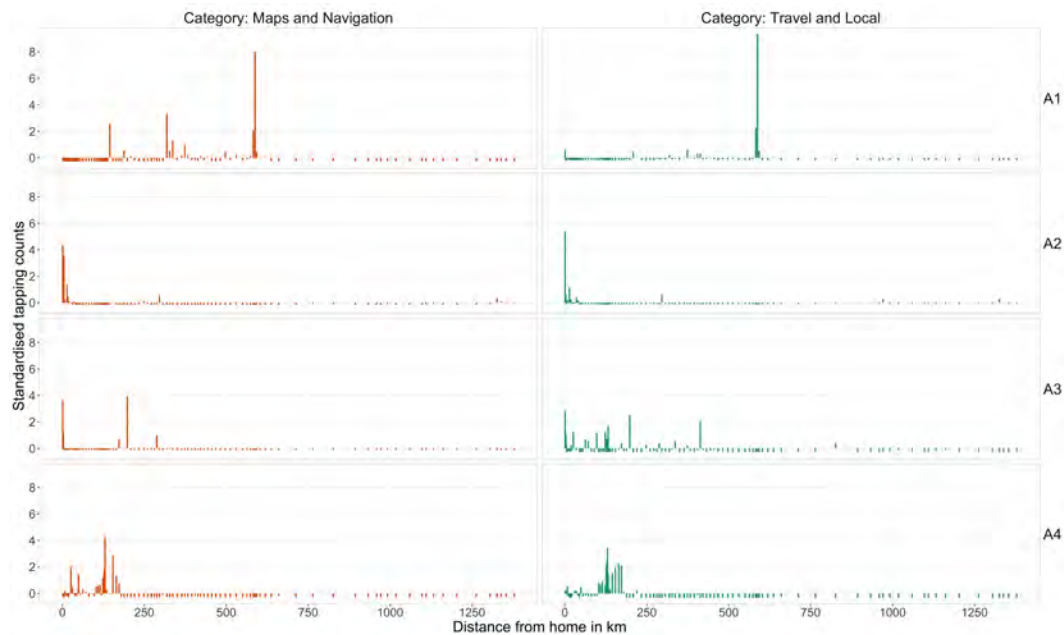
behaviour, with peaks at distances ranging from 5 km to 10 km, 100 km to 150 km, around 300 km and an additional smaller interaction peak at a distance of 1300 km from home. This trend is similar for both categories.

For comparison purposes, we also applied Archetypal Analysis as an unsupervised machine learning technique. Archetypal analysis finds unique combinations of features (or “pure types”) in a dataset (i.e. archetypes) that best represent its properties [3, 2]. Data points of the dataset are then positioned on a spectrum between archetypes without being assigned to only one particular archetype (unlike the results of cluster analysis). With archetypal analysis, we can assess the membership of each data point to these different archetypal signatures (similar to cluster analysis) while preserving individual differences [2]. Based on the RSS value, we chose four archetypes to best represent our data (RSS value of 0.68)

Figure 2 shows the participants’ distribution on the archetypal spectrum of the four calculated archetypes. Most of our participants have strong affiliations to archetype A2, as most data points are around that archetype. The directional lines of each data point indicate the direction and strength of affiliation to the different archetypes. Based on that, most data points near A2 also have strong affiliations to archetype A3. Figure 3 visualises the standardised tapping counts for each archetype over home distances and for each app category. While the signature of each archetype differs in some ways, the tapping behaviour between the two app categories for each archetype is rather similar. Archetype A2 is defined by a strong usage behaviour of both app categories and a distance that is close to home (mainly within a range of 13 km). Many of our participants also have strong affiliations to archetype A3. A3 is defined by a usage behaviour that is mainly distributed over a distance range up to 200 km from home. Comparing archetypes A2 and A3, it is possible to derive differences in interaction behaviour. A2 consists of a behaviour where participants used both app categories and are close to home; A3 indicates a behaviour where participants are also farther away from home, with a scattered and predominant usage behaviour of the TL



■ **Figure 2** Distribution of participants on the archetypal spectrum.



■ **Figure 3** Bar chart displaying the results of the archetypal analysis. Tapping data distribution is plotted for the four archetypes for the selected categories of Maps and Navigation (left) and Travel and Local (right).

app category. Hence, we see a distinction between A2 (home behaviour) and A3 (travel behaviour). Archetypes A1 and A4 show distinct behaviour and could be considered outliers, with one and two participants affiliated with these archetypes, respectively. The tapping signature of A1 and A4 is defined by using both app categories at distances mostly between 300 km and 600 km for A1 and between 100 km and 200 km for A4.

Comparing the cluster analyses results with those of the archetypal analysis, we found two main interaction behaviours: home behaviour and travel behaviour. However, archetypal analysis also allowed us to identify a spectrum of participants' interactions and their direction

towards the different archetypes, which is an advantage over cluster techniques such as HCA. In terms of limitations, we aggregated map-related apps to the category level of each app that the Google Play Store provided. Although we initially aimed to analyse each app's individual tapping data, the recorded apps exhibited high usage variability and frequency among participants. This resulted in scattered contributions from each app, which we considered insufficient for an individual analysis. Future studies should include more participants and collect consistent data points for individual apps to overcome this limitation.

4 Conclusion

This study aimed to expand our understanding of everyday map app usage by extracting as much information as possible from a minimal set of data. Our results provide distinct tapping signatures that point to how participants' app usage behaviour may differ at different distances from home. This is a valuable starting point for evaluating tappigraphy as a method for collecting behavioural data on mobile map use in a non-intrusive and continuous manner. In future studies, we plan to extend our research by using tappigraphy in combination with additional sensors of smartphones (e.g. accelerometer, gyroscope, ambient light sensor, etc.) to consider interactions with map apps in relation to environmental factors.

References

- 1 Myriam Balerna and Arko Ghosh. The details of past actions on a smartphone touchscreen are reflected by intrinsic sensorimotor dynamics. *npj Digital Medicine*, 1:4, 2018. doi:10.1038/s41746-017-0011-3.
- 2 Mona Bartling, Clemens R. Havas, Stefan Wegenkittl, Tumasch Reichenbacher, and Bernd Resch. Modeling Patterns in Map Use Contexts and Mobile Map Design Usability. *ISPRS International Journal of Geo-Information*, 10(8):527, August 2021. Number: 8 Publisher: Multidisciplinary Digital Publishing Institute. doi:10.3390/ijgi10080527.
- 3 Manuel J. A. Eugster and Friedrich Leisch. From Spider-Man to Hero — Archetypal Analysis in R. *Journal of Statistical Software*, 30:1–23, April 2009. doi:10.18637/jss.v030.i08.
- 4 Fionn Murtagh and Pedro Contreras. Algorithms for hierarchical clustering: an overview, II. *WIREs Data Mining and Knowledge Discovery*, 7(6):e1219, 2017. doi:10.1002/widm.1219.
- 5 Tumasch Reichenbacher, Meysam Aliakbarian, Arko Ghosh, and Sara I. Fabrikant. Tappigraphy: continuous ambulatory assessment and analysis of in-situ map app use behaviour. *Journal of Location Based Services*, 16(3), July 2022. Publisher: Taylor & Francis. doi:10.1080/17489725.2022.2105410.
- 6 Robert E. Roth, Arzu Çöltekin, Luciene Delazari, Homero Fonseca Filho, Amy Griffin, Andreas Hall, Jari Korpi, Ismini Lokka, André Mendonça, Kristien Ooms, and Corné P.J.M. van Elzakker. User studies in cartography: opportunities for empirical research on interactive maps and visualizations. *International Journal of Cartography*, 3(sup1), October 2017. Publisher: Taylor & Francis. doi:10.1080/23729333.2017.1288534.
- 7 Gian-Luca Savino, Miriam Sturdee, Simon Rundé, Christine Lohmeier, Brent Hecht, Catia Prandi, Nuno Jardim Nunes, and Johannes Schöning. MapRecorder: analysing real-world usage of mobile map applications. *Behaviour & Information Technology*, 40(7):646–662, 2021. doi:10.1080/0144929X.2020.1714733.
- 8 Johannes Schöning, Michael Rohs, Antonio Krüger, and Christoph Stasch. Improving the Communication of Spatial Information in Crisis Response by Combining Paper Maps and Mobile Devices. In Jobst Löffler and Markus Klann, editors, *Mobile Response*, Lecture Notes in Computer Science, pages 57–65, Berlin, Heidelberg, 2009. Springer. doi:10.1007/978-3-642-00440-7_6.