Applied Geography 158 (2023) 102997

Contents lists available at ScienceDirect

Applied Geography

journal homepage: www.elsevier.com/locate/apgeog

Assessing the socio-demographic representativeness of mobile phone application data

Michael Sinclair [a,*], Saeed Maadi [a], Qunshan Zhao [a], Jinhyun Hong [b], Andrea Ghermandi [c], Nick Bailey [a]

[a] *Urban Big Data Centre, University of Glasgow, Glasgow, UK*
[b] *Department of Smart Cities, University of Seoul, South Korea*
[c] *Department of Natural Resources and Environmental Management, University of Haifa, Israel*

ARTICLE INFO

Handling Editor: Y.D. Wei

Keywords:
Mobile phone data
Socio-demographic representativeness
Tamoco
Huq

ABSTRACT

Emerging forms of mobile phone data generated from the use of mobile phone applications have the potential to advance scientific research across a range of disciplines. However, there are risks regarding uncertainties in the socio-demographic representativeness of these data, which may introduce bias and mislead policy recommendations. This paper addresses the issue directly by developing a novel approach to assessing socio-demographic representativeness, demonstrating this with two large independent mobile phone application datasets, Huq and Tamoco, each with three years data for a large and diverse city-region (Glasgow, Scotland) home to over 1.8 million people. We advance methods for detecting home location by including high-resolution land use data in the process and test representativeness across multiple dimensions. Our findings offer greater confidence in using mobile phone app data for research and planning. Both datasets show good representativeness compared to the known population distribution. Indeed, they achieve better population coverage than the 'gold standard' random sample survey which is the alternative source of data on population mobility in this region. More importantly, our approach provides an improved benchmark for assessing the quality of similar data sources in the future.

1. Introduction

New forms of mobile phone (MP) data from the use of applications, or 'apps', offer enormous potential as an alternative or complement to traditional survey data sources to enhance our understanding of human activity and mobility (Huang et al., 2022; Kang et al., 2020). The huge volumes of data available from these novel sources as well as the spatial and temporal details they provide, create unprecedented opportunities across a wide range of disciplines to advance scientific research. However, there are critical unanswered questions concerning the socio-demographic representativeness of these new forms of MP data. This creates a risk that underlying bias could produce unreliable results which are then used as the basis for policy (Grantz et al., 2020). Furthermore, the limited analysis of the issue of socio-demographic representativeness restricts the progress of applied research seeking to utilise these novel and emerging form of MP app data as an alternative or complement to more traditional data sources.

Traditionally, scientific research has utilized MP data from call detail records, which track phone locations during potentially billable events (Calabrese et al., 2013; Grantz et al., 2020; Pappalardo et al., 2021; Ren & Guan, 2022; Vanhoof et al., 2018b; Wang et al., 2020; Yabe et al., 2022). More recently, a new form of location data from the use of GPS-enabled smartphone applications has emerged, which also offer large data volumes but with much higher spatial accuracy (Berke et al., 2022; Grantz et al., 2020; Huang et al., 2022; Wang et al., 2020; Yabe et al., 2020). This mobile phone application (MPA) data, generated and collected from the use of a wide range of apps, provides point location information which supports more detailed analysis and opens up the range of possible analytical applications (Cameron et al., 2020; Heo et al., 2020; Mears et al., 2021; Sinclair et al., 2021; Yabe et al., 2020). So far, these have included disaster and pandemic response (Huang et al., 2022; Kishore et al., 2022; Yabe et al., 2020), nature-based recreation (Mears et al., 2021; Sinclair et al., 2021) and analyses of human mobility (Calafiore et al., 2021; Gao et al., 2020; Kang et al., 2020). The applications of these novel data are in their infancy and their potential spans a wide range of disciplines.

[*] Corresponding author.
E-mail address: michael.sinclair@glasgow.ac.uk (M. Sinclair).

https://doi.org/10.1016/j.apgeog.2023.102997
Received 17 February 2023; Received in revised form 20 April 2023; Accepted 11 May 2023
Available online 13 July 2023

New spatial data forms such as MPA data are frequently contrasted with traditional survey data as an alternative source for applied research (Mayer-Schönberger and Cukier, 2013; Savage & Burrows, 2007). Although household surveys are considered the 'gold standard' for research due to generalizable random samples, they face declining response rates (Brick & Williams, 2013; Meyer et al., 2015), interviewer effects, recall error and normative bias (Marsh, 1982). Surveys are unsuitable for rapid response situations like the Covid-19 pandemic, and their low sample sizes limit fine-grained spatial/temporal detail. Additionally, some highly marginalized groups such as the homeless or those in temporary forms of accommodation may be excluded. In this context, MPA data may offer advantages as a complement or alternative to traditional data (boyd & Crawford, 2012). MPA data provide spatial-temporal detail, often in (near) real-time or with low lag. While consent is still required, data collection is less burdensome, potentially reducing non-response bias and including previously excluded groups. Despite this potential, there are uncertainties of these novel data. A major one concerns data quality since so much of the data capture and processing is unavailable to researchers due to commercial concerns. This raises worries that the data may under-represent marginalized groups, particularly the 'digitally excluded' (boyd & Crawford, 2012).

There are critical questions, therefore, about the quality of MPA data which need to be addressed before more widespread use is justified in research (Grantz et al., 2020). Key among them is the question of how accurately MPA data represents the population of interest, given they are not the result of a carefully-planned sampling strategy (Ranjan et al., 2012; Zhao et al., 2016) and are constructed from the use a wide range of different applications. In particular, the concern is that inequalities in access to and use of mobile phones may be reflected in these data. The resulting research may skew attention and possibly resources towards already advantaged social groups (Grantz et al., 2020) or fail to adequately include groups such as the elderly population (Guo et al., 2019; Lee et al., 2021). Though the question of bias is common to all forms of MP data, it is perhaps especially relevant to MPA data where datasets are assembled by commercial intermediaries. These intermediaries gather data across a wide and diverse set of apps with the aim of achieving scale and broad representativeness, but these are not transparent with few metrics provided to evidence the latter and there are currently no standards by which they can be evaluated. Despite its importance, few studies using MPA data explore the topic of socio-demographic representativeness directly (Huang et al., 2020, 2022).

Assessing representativeness is challenging due to the steps taken by MP data providers to protect user privacy. MP data are often provided to researchers as aggregated totals, making it impossible to identify the characteristics of individuals at all. Where data are provided at the individual level, such as with MPA data, information is rarely if ever provided on a user's personal characteristics, so representativeness cannot be examined directly (Grantz et al., 2020). Researchers therefore often use techniques to infer the user's home location based on location histories. These home locations allow the geographic distribution of the sample of MP users to be compared to 'ground truth' sources such as official population statistics (Berke et al., 2022; Calabrese et al., 2013; Huang et al., 2022; Mao et al., 2015; Phithakkitnukoon et al., 2012; Wang et al., 2019; Çolak et al., 2015). This process provides a very useful measure of differential geographic coverage (Yabe et al., 2020) as well as variations by the socio-demographic status of different areas (Bernabeu-Bautista et al., 2021; Huang et al., 2020, 2020, 2022, 2020). Enriching the data in this way also greatly increases the potential impact of research.

Different approaches have been adopted to estimate or infer home locations from MP and other locational data based on the volume of content generated by a user in space and/or time (Calafiore et al., 2021; Pappalardo et al., 2021; Sinclair et al., 2020). Since it is rarely possible to validate home detection algorithms against known home locations for users (Pappalardo et al., 2021), techniques are designed with the aim of reducing potential error. To assign a home location at a country or city level, daily activity counts are generally sufficient (Bojic et al., 2015; Sinclair et al., 2020). However, to infer socio-demographic information for users requires predicting home locations for much smaller geographies. Including the full range of an individual's daily activities towards this end could lead to an increase in false predictions as users might record volumes of data around places designated for work or socialising (Pappalardo et al., 2021; Vanhoof et al., 2018a). This is especially true for MPA data where datasets represent a wide range of activities, due to the mix of apps involved. The main approach to overcome this is to utilise activity heuristics, by including a time element in the algorithm. Restricting the analysis to night-time data, based on the assumption that people are more often at home during this period (Berke et al., 2022; Bojic et al., 2015; Calabrese et al., 2013; Calafiore et al., 2021; Phithakkitnukoon et al., 2012; Sinclair et al., 2020; Vanhoof et al., 2018; Çolak et al., 2015), has been shown to improve results (Pappalardo et al., 2021).

There are two main limitations with current approaches to assessing representativeness of these novel data. The first is that, even with activity heuristics, problems remain in inferring home locations as many people spend periods of the night at sites of work, leisure, or transit. This is especially true for MPA data where the data represent various activities based on a diversity of apps. The second is that, once home locations have been inferred, researchers rarely explore representativeness in a systematic or comprehensive way. In this paper, we address both issues and hence provide a more appropriate standard for assessing representativeness of MPA data. First, with home locations, we propose a novel approach which incorporates high-resolution land use data into the process. By relying only on data captured within buildings which have a designated residential use, we greatly reduce the chance of identifying night-time work, leisure, or transit locations as home locations. Second, we use these potentially improved home location estimates to examine representativeness using multiple independent dimensions. These cover the geographical distribution but also socio-economic and socio-demographic status.

To illustrate our approach, we apply it to an assessment of representativeness for two extensive and independent sources of MPA data. Each contains data from a diverse portfolio of apps covering a wide time period (three years) for a large and socio-demographically diverse city-region (Glasgow). First, we apply our home detection approach which incorporates high-resolution residential land use data into the process. Second, we compare the distribution of the resulting samples of MPA users from both data sources to the known population distribution across three years (2019–2021). Comparisons are made by geographic location and against two different measures of area socio-demographic status. One is an official index of area deprivation in Scotland, the Scottish Index of Multiple Deprivation (SIMD). The other is a commercial socio-demographic classification, CACI's Acorn consumer classification (CACI), which segments areas by analysing a wide range of data on demographics and consumer behaviour. Third, we compare the results on representativeness found using our novel home detection approach against those found using a more conventional approach, which does not utilise residential land use data, to illustrate the impact of this innovation. Finally, we compare the distribution of our sample of MP users to the distribution of the sample of households captured by a traditional survey which is widely used in the study area for mobility analysis and transport planning, the Scottish Household Survey (SHS). Such traditional forms of data are frequently held up as the 'gold standard' against which new forms of data are compared since they are built round a structured random sample. Comparison against such a sample provides arguably a fairer test of representativeness.

2. Data and methods

2.1. Study area

The study area is Glasgow city-region, comprising the core city (the largest in Scotland) and seven surrounding councils (Fig. 1). Glasgow is home to over 600,000 people, while the wider city-region houses over 1.8 million people. The city-region covers areas or neighbourhoods with a wide range of socio-demographic circumstances, which makes it particularly suitable to test for inequalities in sample coverage by socio-economic status. Fig. 1 also shows the eight council boundaries used for reporting results as well as the built-up residential areas within each.

2.2. Mobile phone application datasets

The core data for this research are MPA datasets from two private companies, Huq and Tamoco[1]. Both are examples of smartphone GPS location data (Yabe et al., 2022) which are timestamped point data generated using MP apps on GPS-enabled smartphones (Table 1). This type of big data generally offers a higher spatial precision than traditional sources of MP data such as call detail records which are often limited to cell tower regions. The data used in this study are confined to the extent of the study area (Fig. 1) and consist of hundreds of millions of data points per year. Wider, Huq currently offers data across the UK and Tamoco across the UK and the United States of America.

The construction and structure of the datasets are similar across both providers. Each contains data from a range of partner apps on an informed consent basis, with data limited to users aged 16+. Data is collected when an app records the time and location of a device based on the most accurate location sensor available at the time, including GPS, Bluetooth, cellular tower, Wi-Fi or a combination of sources (Wang & Chen, 2018). Due to a lack of transparency from the commercial providers, the specific applications included in the datasets are unknown to researchers. However, data are pooled from a wide and diverse set of apps with the aim of achieving scale and broad representativeness. In one of the years, for example, one provider was collecting data from over 200 unique apps. The data represent timestamped point locations with a certain degree of error. Each MP device has the personal identifiers replaced with non-reversible hashed identifiers. This means that data points from the same user can be linked over time. With Huq, the points from individual users can be linked over the whole period while Tamoco resets its hashed identifiers every month. Data volumes are vast and fluctuate year-to-year (Table 1), reflecting in part the changes in the apps with whom the intermediaries have contracts. The challenge of assessing representativeness will therefore always be an on-going exercise.

2.3. Other secondary data sources used in the analysis

Different levels of geographic boundaries are used in the analysis, all of which are represented visually in Appendix 1. The highest level used is Council (n = 8) which is also visualised in Fig. 1. The next level is the Intermediate Zone (n = 417) which nest within councils. We also use Datazones, which nest within Intermediate Zones, and are the key geography for small area statistics in Scotland. These are the spatial unit used in this paper for home location detection. Datazones are also used to assign the Scottish Index of Multiple Deprivation measure of socio-demographic status to mobile devices (see below). Datazones are designed to have a population of 500–1000 and there are 2336 in the study area. The finest spatial boundary used is the unit postcode (n = 44,829) which nest within Datazones. These boundaries are used to assign the second measure of socio-demographic status to users, the CACI Acorn Consumer Classification (see below).

In comparing the socio-demographic representativeness of MP users to the population, we use two sources, one public and one private. The Scottish Index of Multiple Deprivation (SIMD, https://simd.scot/) assigns a relative measure of area deprivation to Datazones across Scotland. The SIMD combines measures of deprivation across multiple domains (income, employment, education, health, crime, housing and access to services) into a scaleless relative ranking. SIMD 2020 is used in this research. Our analysis assigns MP users an SIMD quintile and percentile, using national rankings, based on the Datazone where they are estimated to live. As Glasgow city-region is a relatively deprived area, there is an over-representation in more deprived quintiles (1 and 2). See the supplementary material for population breakdown by SIMD groups. The CACI Acorn Consumer Classification (https://www.caci.co.uk/) is a private socio-demographic data source which segments the UK population by analysing a wide range of data on demographics and consumer behaviour. CACI segments unit postcodes into 6 categories, 18 groups and 62 types. The subdivisions are nested, with the 6 categories broken into between 2 and 4 groups[2], and the 18 groups broken into between 3 and 6 types. Our analysis assigns mobile phone users with a category, group and type based on the postcode where they are estimated to live. In this study we use 2020 CACI data. See the supplementary material for population breakdown by CACI groups.

As a key step in the home location detection process, which is explained in the next section, we make use of high-resolution land use data from Geomni's UKBuildings layer. This dataset is a multi-polygon spatial dataset representing the footprint of all buildings in the UK, including residential buildings (see Fig. 1). Each building is assigned a usage, classified into various types. For this study, we use all buildings with a residential or mixed-residential use. Data from 2020 is used in this study.

In the final section of the results, we compare the MPA samples to a traditional survey dataset widely used in social research across Scotland, the Scottish Household Survey (SHS, http://www.scottishhouseholdsurvey.com/). The SHS is an annual survey of over 10,000 households, used as the basis of a range of official statistics. The SHS has a repeat cross-sectional design with a sample for the Glasgow city-region of N = 3495 in 2019 (the most recent available).[3] For Glasgow City, less than half the eligible adults completed a travel diary. Younger adults were significantly under-represented while those 65+ were over-represented. Some population groups are excluded by the sample design including households living on military bases, in communal establishments, in mobile homes or sites for traveling people, or homeless (Scottish Government, 2020).

2.4. Home location detection techniques

To compare the distribution of each MPA sample to the population, it is necessary to estimate the home location for each MP user in the dataset and this is typically done based on night-time locations, as discussed in the Introduction. The specific period which constitutes night-time varies between studies but a window beginning between 19.00 and

[1] Information for Huq available at: https://www.ubdc.ac.uk/data-services/data-catalogue/transport-and-mobility-data/huq-data/; and Tamoco available at: https://www.ubdc.ac.uk/data-services/data-catalogue/transport-and-mobility-data/tamoco-data/.

[2] This is with the exception of the category/group 'Not Private Households' which is not disaggregated between the levels of category and group (and related to areas which generally do not have a residential population).

[3] Households are selected using a random sample stratified by council which over-represents smaller councils to ensure each achieves a minimum sample size. A travel survey portion is completed by one randomly-selected adult in each household (Scottish Government, 2020) and is by definition therefore skewed towards adults from smaller households. For 2019, the response rate for households nationally was 63%, with random adults completing the travel survey in 92% of cases but this varied across the country.

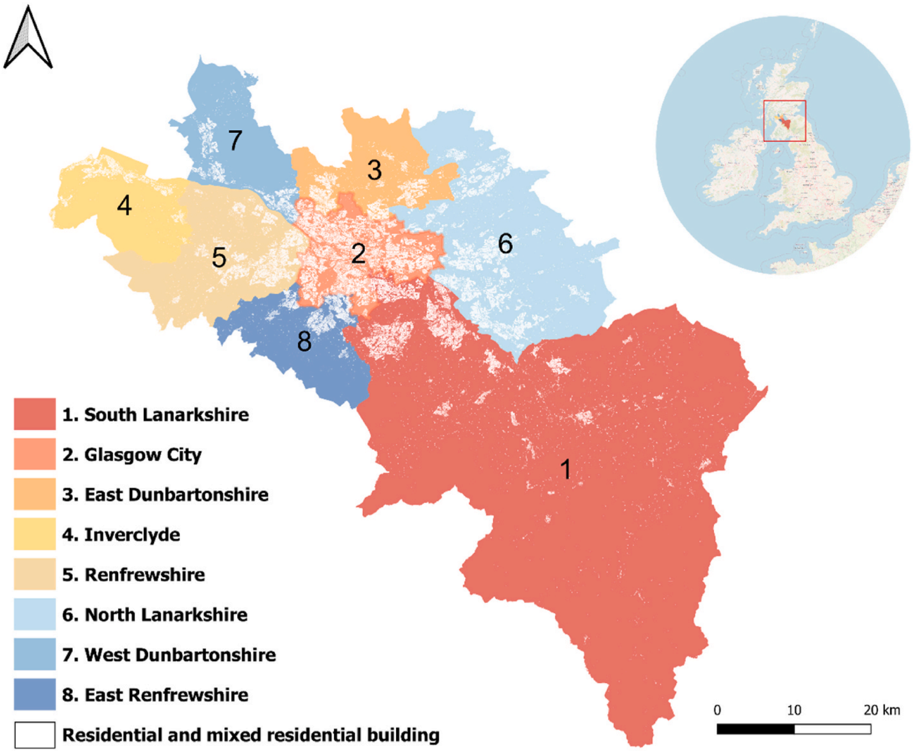


Fig. 1. Glasgow City-region, council areas and built-up areas
Residential and mixed residential buildings are from Geomni's UKBuildings layer which is created and maintained by Geomni, a Verisk company (see section on data sources).

Table 1
Summary of mobile phone application data collections in the study area.

Provider	Measure	2019	2020	2021
Huq	Unique users	19,399	29,741	25,233
	Datapoints (millions)	21.9	161.8	346.8
	Mean datapoints per user	1129	5440	13,744
Tamoco	Unique users	81,203	85,258	81,136
	Datapoints (millions)	442.5	808.1	471.8
	Mean datapoints per user	5449	9478	5814

Notes: Unique users are based on the number of unique hashed identifiers active in a given year. The number of Tamoco users is based on the monthly average for each year since identifiers are reset monthly.

22.00 and ending between 05.00 and 09.00 is common (Pappalardo et al., 2021; Vanhoof et al., 2018). It is rarely possible to verify the accuracy of home location estimates with 'ground truth' data. One study which achieved this found that using night-time data was more accurate than taking data which covered the whole day (Pappalardo et al., 2021). Accordingly, this is the approach we build on here using the night time period of 20.00 to 06.00. Box 1 explains in more details how we estimate home locations using our approach (Method 1) and a more conventional approach (Method 2).

2.5. Comparing the representativeness of mobile phone application data

The results from section 2.4 allow us to allocate each unique MP user to a Council, Intermediate Zone, Datazone, and unit postcode. Using these we can assign MP users to a SIMD deprivation quintile and percentile (from Datazone), as well as a CACI Acorn category, group, and type (from postcode). We assess representativeness in three ways. For the geographic distribution, we focus initially on the eight council areas but later report the distribution across the 417 Intermediate Zones. For variations by deprivation status, we initially examine the distribution across the five quintiles of the SIMD index but later present results at the percentile level. Lastly, for variations by socio-demographic status, we initially examine the distribution across CACI's six broadest categories but later use the 18 groups and the 62 types. Following these comparison to the population, we make a further comparison with the sample of travel diary respondents in the SHS. We make comparisons across the eight councils using the measure of SIMD deprivation quintile, the latter being the finest spatial disaggregation available on the publicly-available SHS files.

2.6. Transparency and reproducibility

The MP datasets used in this analysis can be accessed for research purposes by application to the Urban Big Data Centre, an Economic and Social Research Council funded research centre and national data service based at the University of Glasgow. Datazone and higher geographic boundaries are available under Open Government license (http://spatialdata.gov.scot/). Postcode boundaries are freely available from the Scottish Postcode Directory (National Records of Scotland, n.d.) under the 'Public Sector Geospatial Agreement' which covers non-commercial use of the data. SIMD data is available under Open Government licence. CACI data are accessed here under a licence agreed with CACI for this particular study. Geomni's UKBuildings layer (Digital Map Data © The GeoInformation Group Limited (2022), created and maintained by Geomni, a Verisk company) is accessed under a general academic license via Digimap (https://digimap.edina.ac.uk/). SHS data are accessed through the UK Data Service under their standard End User Licence (Scottish Government & Ipsos MORI, 2021). All analysis is completed using a combination of PostgreSQL and R programming language (R Core Team, 2022). The code to process the data and estimate home location is openly available on GitHub (https://github.com/sinclairmichael/appliedgeography_representativeness.git).

Box 1
Summary of home location detection techniques

Method 1: Using activity heuristics and land use to estimate home location.

Each user's datapoints (see Table 1) are first limited to those with an accuracy reported as 100m or better. Their data is then further limited to that which falls within the footprint of a building identified as having residential or mixed-residential use. Using this subset, the home location for each user is estimated as the Datazone where they record the maximum number of active evenings in the study area during the time period[41], where an evening is considered between 20.00 and 06.00 h. For Huq, the home location is estimated using data across each calendar year. For Tamoco, home locations have to be estimated monthly because the user ID is re-hashed each month (see section on mobile data sources). To identify a unit postcode to the home location (which is important to assign the private socio-demographic data to the user), we take the set of night-time residential datapoints within the home Datazone and identify the postcode which contains the average (geographic centroid).

Method 2: Using activity heuristics only to estimate home location.

As with Method 1, each user's datapoints are limited to those with an accuracy reported as 100m or better. This method does not apply the additional restriction of being within a building of residential or mixed-residential use. For each unique user, as previously, the home location is the Datazone where they record the most active evenings in the study area during the time period[4], where an evening is considered between 20.00 and 06.00 h. Home locations are estimated yearly for Huq and monthly for Tamoco as with Method 1. A unit postcode is similarly assigned as with Method 1 based on the average (geographic centroid) of the night-time datapoints within the home Datazone.

3. Results

3.1. Estimated home locations of mobile phone application data

Results on home location detection for both methods are presented in Table 2. Applying the more restrictive home detection approach which incorporates residential land use data (Method 1), we allocate between 20% and 37% of users with a home location across the years. These users are responsible for generating over 75% of the annual data in all but one case (Tamoco in 2019). The more basic and less restrictive home detection approach (Method 2) leads to a large number of home location estimates of between 37% and 51%. These users are responsible for generating between 88% and 98% of the data in a given year. In both cases, the algorithms tend to cut out the long 'tail' of users who generate relatively few datapoints. These users may be relatively inactive or make infrequent use of the relevant apps, or they may be occasional visitors from outside the region. Overall, a greater portion of Huq users are assigned a home location than Tamoco. This may be due to the persistent hashed identifier for Huq which provides a picture of movements for each user over a wider time period whereas data for Tamoco users can only be linked over one month.

3.2. Representativeness of mobile phone application data

Using our home detection approach which includes residential land use data (Method 1), Figs. 2 and 3 show three measures of sample representativeness for the Huq and Tamoco datasets respectively. The Figures show the distribution geographically by council area (N = 8), by deprivation quintile (N = 5) and by broad CACI Acorn group (N = 6). Overall, the Figures show a very close fit between the distribution of both samples and the population on each measure. There is some variation from year to year but also a great deal of consistency. In terms of council areas, the sample share is almost always within 2.5% of the population share. The exceptions are for Glasgow City where there is some under-representation across both datasets in one or two years, although proportionately this is still a modest deviation given Glasgow is home to more than a third of the city-region population.

A concern that these new forms of data may fail to adequately capture poorer groups does not appear justified. With both datasets, there is both under- and over-representation of the most deprived area (quintile 1) of the SIMD with the same true of the 'Urban Adversity' category from CACI, which will cover similar population groups. Where the Tamoco sample shows a shift towards more affluent areas over time, the opposite trend is observed with Huq. Both MPA datasets have a very small number of users in locations identified by CACI as 'Not Private Households'. These are locations with primarily non-residential uses such as retail, industry, or transport infrastructure. Despite the label, they are home to 1.1% of the population and a similar proportion of the two samples.

Using the CACI classification, we can make comparisons at a finer level based on 18 groups, focusing on 2020 only for clarity where there is a direct comparison to population data (Fig. 4). As before, we find a very good fit to the population across both data sources. While we might be concerned in advance about risks of under-representation of less affluent groups and over-representation of more affluent groups, there is little evidence from our analysis to support this, with the possible

Table 2
Number of users assigned home location by algorithms without and with land use.

Metric	Dataset	Users and datapoints assigned with home location					
		Method 1: Using activity heuristics and land use			Method 2: Using activity heuristics only		
		2019	2019	2019	2019	2020	2021
Unique users	Huq	4,633 (24.0%)	11,079 (37.3%)	9,165 (36.3%)	7,204 (37.1%)	15,135 (50.9%)	12,495 (49.5%)
	Tamoco	21,031 (25.9%)	20,355 (23.9%)	16,586 (20.4%)	36,025 (44.4%)	34,693 (40.7%)	29,682 (36.6%)
Datapoints (millions)	Huq	18.8 (85.7%)	151.7 (93.7%)	331.0 (95.4%)	20.9 (91.8%)	161.8 (97.7%)	346.8 (98.4%)
	Tamoco	253.1 (57.2%)	659.9 (81.7%)	366.8 (77.8%)	389.3 (88.0%)	749.0 (92.7%)	410.3 (87.0%)

The number of unique users for Tamoco is based on the monthly average for each year since identifiers are reset each month. Values in parentheses are percentages of the totals shown in Table 1. See Methods section for further details.

exception of the Tamoco coverage of the groups 'Executive Wealth' and 'Difficult Circumstance'. On the contrary, the groups 'Modest Means' and 'Striving Families' are slightly over-represented in both datasets, though differences are modest. Another group where we have particular concerns about under-representation in MP data is in the elderly population but we do not find evidence of that here. The proportions of our sample in 'Mature Money' or 'Comfortable Seniors' groups is very close to that for the population in both datasets. With areas home to groups both older and poorer ('Poorer Pensioners'), we do find these under-represented by both datasets but only slightly; they make up 8.6% of Huq users and 8.8% of Tamoco users compared with 9.9% of the population.

Taking the geographic and socio-demographic comparison to a finer level of disaggregation still, we switch to summarising results using the correlation between the estimated number of MPA users in an area and the (adult) population of that area. Table 3 (top panel) shows results for each dataset for each year, and for: sub-council area geographies (417 Intermediate Zones); percentiles of deprivation on the SIMD (100 groups); and the CACI Acorn types (62 groups). The correlations for each level are very similar across years and between datasets. For the geographic comparison across Intermediate Zones, there are moderate to strong correlations for both Huq (0.58–0.63) and Tamoco (0.57–0.69). For the comparisons by socio-demographic status, we see very strong correlations across both datasets and all three years of analysis with only one correlation below 0.9 (for Huq in 2019 when it had a far smaller sample size).

Table 3 (lower panel) takes the analysis a stage further, examining correlations for the two socio-demographic measures within individual councils (n = 8), for each dataset and each year. Again, the picture is of a high degree of representativeness. In 2020, the correlations with SIMD percentiles within councils range between 0.82 and 0.97 for Huq and 0.87–0.98 for Tamoco. The correlations with CACI types within councils range between 0.95 and 0.99 for Huq and 0.93–0.98 for Tamoco. For both datasets, the highest correlations are found in the largest council, Glasgow City, while the lowest are found in the smallest, Inverclyde.

3.3. *Comparing home detection algorithms*

We compare our approach (Method 1) with the basic home detection algorithm (Method 2) by looking at how the sample distributions from each approach compare with the population distribution. One clear point of difference is in relation to the proportion of each sample estimated to live in locations identified by CACI as 'Not Private Households' (i.e., places where the dominant land use is not residential). Method 2 estimates that 6.3% of Huq users and 5.9% of Tamoco users lived in these locations in 2020, compared with just 1.1% of the population, presumably because people are working, socialising or traveling in these locations over a number of evenings and misattributed to live there. Method 1, by restricting the data to points within buildings with a residential use, obtains a proportion much closer to the expected proportion (1.5% and 1.0% respectively), as we discuss in the previous subsection.

More broadly, we can examine how correlations between the samples and the population change when moving from the basic (Method 2) to the refined approach (Method 1), again focusing on 2020 (Fig. 5). Here we show *changes* in the correlation between samples and the population comparing Method 2 with Method 1. We show this for Intermediate Zones (geographic), percentiles (SIMD), and types (CACI). A *positive difference* indicates that Method 1 (our approach) yields a *higher* correlation, i.e., a sample distribution more similar to that of the population. A higher correlation does not, of course, prove that the more restrictive approach is more accurate at identifying the true home locations. We argue, however, that a closer correlation, combined with the specific improvement in relation to 'Not Private Household' locations, is strongly suggestive of a better accuracy. The figures show a higher correlation in every case when using our approach (Method 1). In general, the increases in correlations are greater with the Huq dataset. It is not clear why this should be the case although it may reflect the mix of apps from which each company is gathering data; Huq may draw on a larger proportion related to transport or mobility, for example, leading the basic algorithm to mis-allocate a greater proportion.

3.4. *Comparing mobile phone app samples with household survey samples*

To assess the potential bias in MPA data relative to traditional survey data, we compare the distribution of the two MPA samples with the distribution of the sample for the main Scottish survey which captures trip data, the SHS (Fig. 6). More details on this survey can be found in section 2.3. Neither sample is weighted here. Overall, the MPA samples have a better coverage of the adult population than the (unweighted) SHS sample, and with a much larger sample size. The mean absolute error of SIMD quintile groups across the eight councils is 1.8% for Huq and 1.3% for Tamoco compared with 2.1% for the SHS. One or both MPA dataset outperforms the SHS in seven of the eight councils. West Dunbartonshire is the only council where the SHS is the most representative although the difference is marginal. Unlike the SHS, MPA data are not skewed to smaller councils nor to smaller households and, while we cannot quantify any bias in the MPA data by age, comparisons across the CACI categories in the previous section (Fig. 4) suggests we do not have such an underlying inequality as the SHS. With the SHS, as with other household survey data where characteristics such as age and sex of respondents are known, weights can be applied to produce a sample which resembles the Scottish population's age-sex distribution by council although this does not of course remove all bias from uneven response rates. With the MPA data, we can only apply weights by geographic area (e.g., council) and/or by socio-demographic status of areas (e.g., SIMD quintile). Based on the results here, however, the weights for MPA data may only need to make small adjustments. Of course, survey data have many other strengths, containing personal characteristics along with information on trip purposes and modes, for example. Nevertheless, the combination of greater representativeness with scale and their spatial and temporal detail gives these MPA data great advantages for many applications.

4. Discussion

This research offers a novel home location detection methodology for MPA data and demonstrates a more comprehensive approach to assessing representativeness across multiple socio-demographic dimensions and spatial scales. We demonstrate the value of our approach in assessing the socio-demographic representativeness of major MPA data collections from two independent providers over an extensive time period and for a major city-region. Our advancement in home location detection using residential land use information improves the fit across both data providers over all three years at the city-region level, and with little loss in data volumes. Using public and private measures of socio-demographic status, we find a good fit between two MPA datasets and the population. While findings are specific to the datasets examined and the spatio-temporal context, our approach provides a valuable benchmark for future assessments as well as further strong encouragement for the use of MPA data in social science research and policy applications.

As Grantz et al. (2020) note, researchers must pay constant attention to the potential biases in MP data from any source. Given that these data are necessarily de-identified to protect privacy, it will always be difficult to tackle the question of representativeness at the individual level. Limited information may be sought on individual user characteristics

[4] Home locations are only assigned when users record two or more active evenings in the Datazones in that period. In cases where more than one potential home location is returned with an equal number of active evenings, the user is not assigned a home location and removed from further analysis.

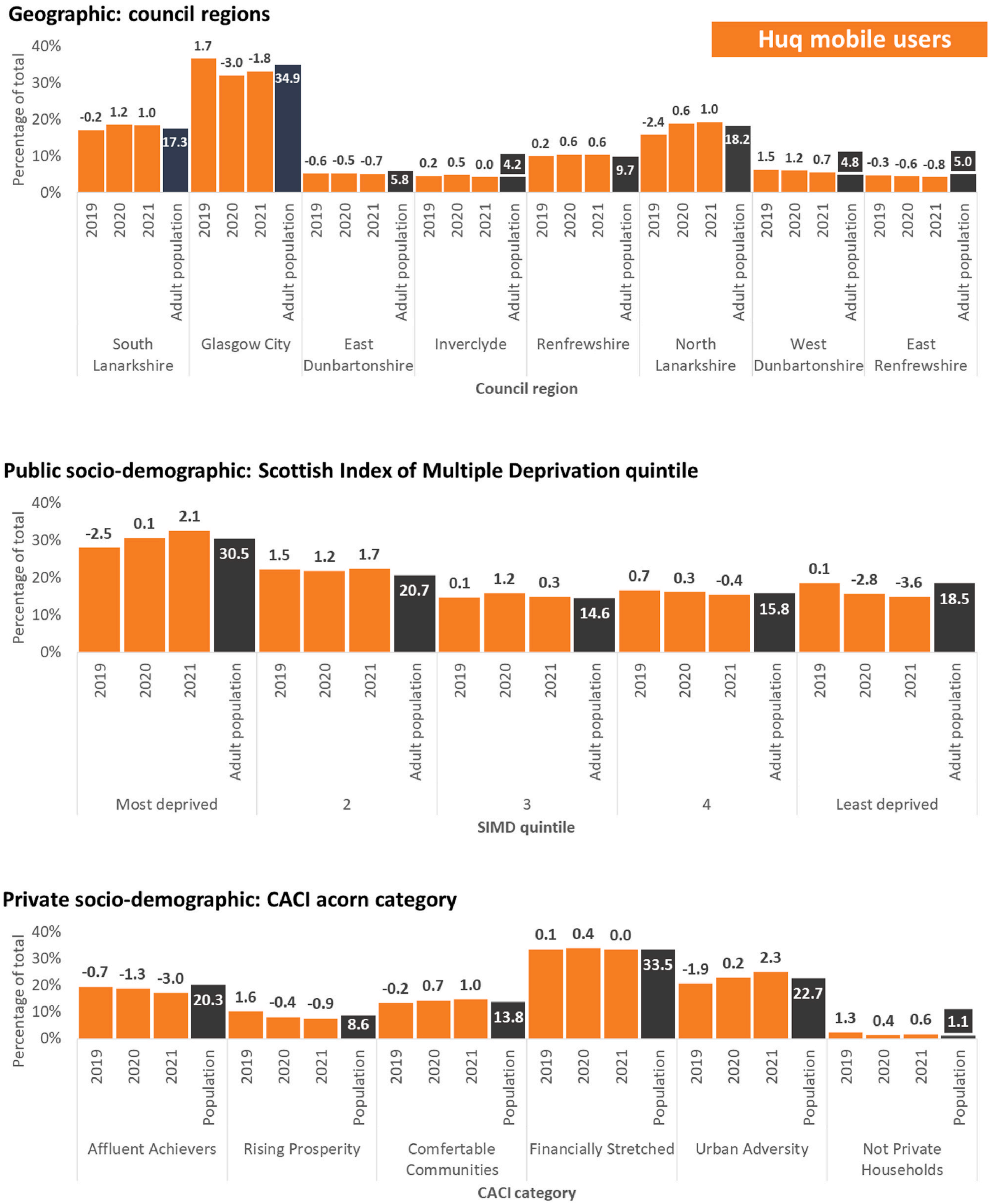


Fig. 2. Geographic and socio-demographic comparison of the Huq mobile population across the Glasgow City-region based on the activity heuristics with land use home detection method.

Colored bars represent the proportion of MPA users in a given year in a category while black bars show the benchmark population. Geographic and public socio-demographic results are based on the adult population in 2020 while the private socio-demographic comparison is based on the total population in 2020 (adult population data is not available at the level used for CACI data). Labels are percentage values (population share or, for MPA data, deviation from this). Quintiles are based on national data and Glasgow city-region has an over-representation of more deprived Datazones. See appendices S1-6 for details on SIMD/CACI population data.

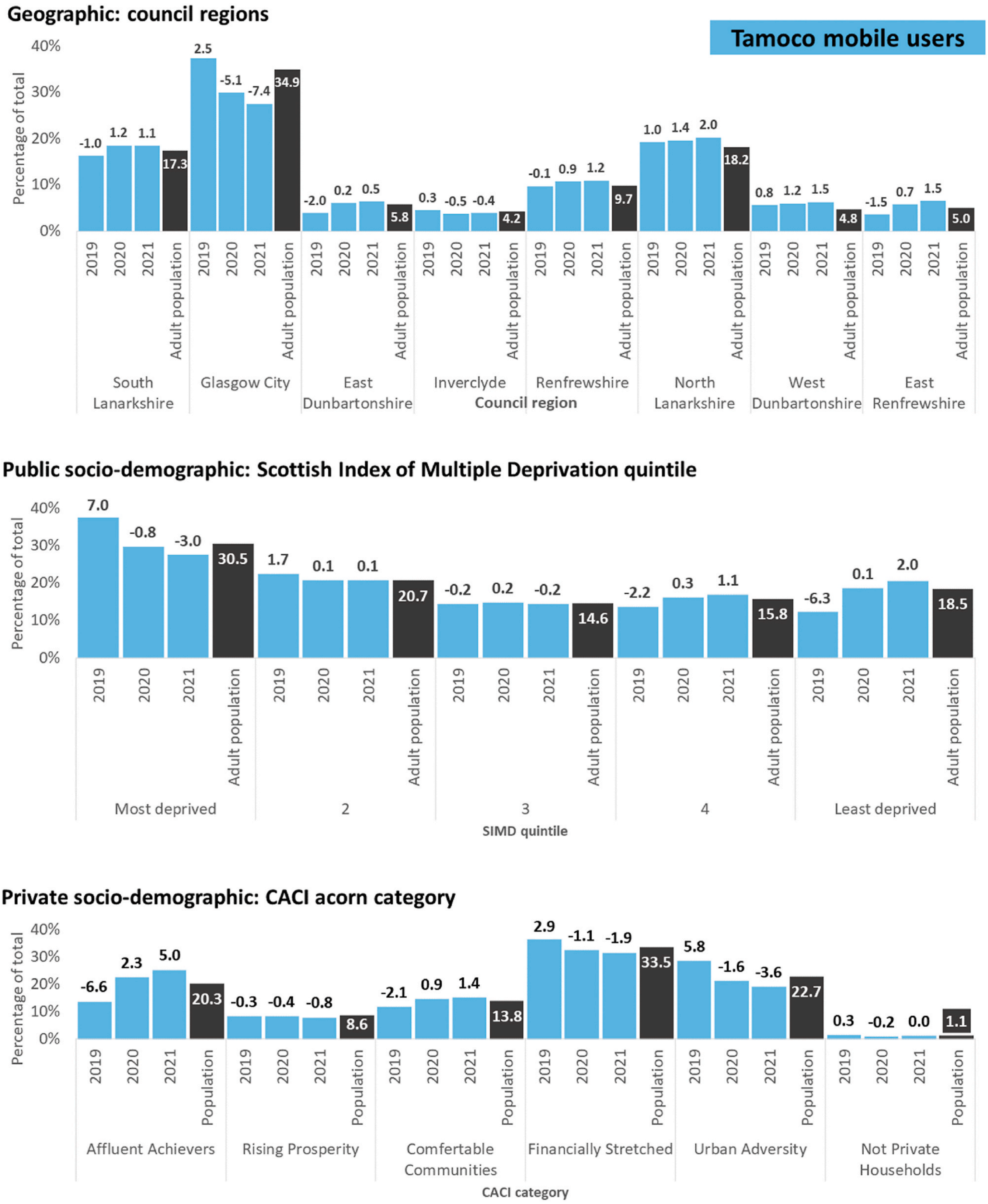


Fig. 3. Geographic and socio-demographic comparison of the Tamoco mobile population across the Glasgow City-region based on the activity heuristics with land use home detection method.

Colored bars represent the proportion of MPA users in a given year in a category while black bars show the benchmark population. Geographic and public socio-demographic results are based on the adult population in 2020 while the private socio-demographic comparison is based on the total population in 2020 (adult population data is not available at the level used for CACI data). Labels are percentage values (population share or, for MPA data, deviation from this). See appendices S1-6 for details on SIMD/CACI population data.

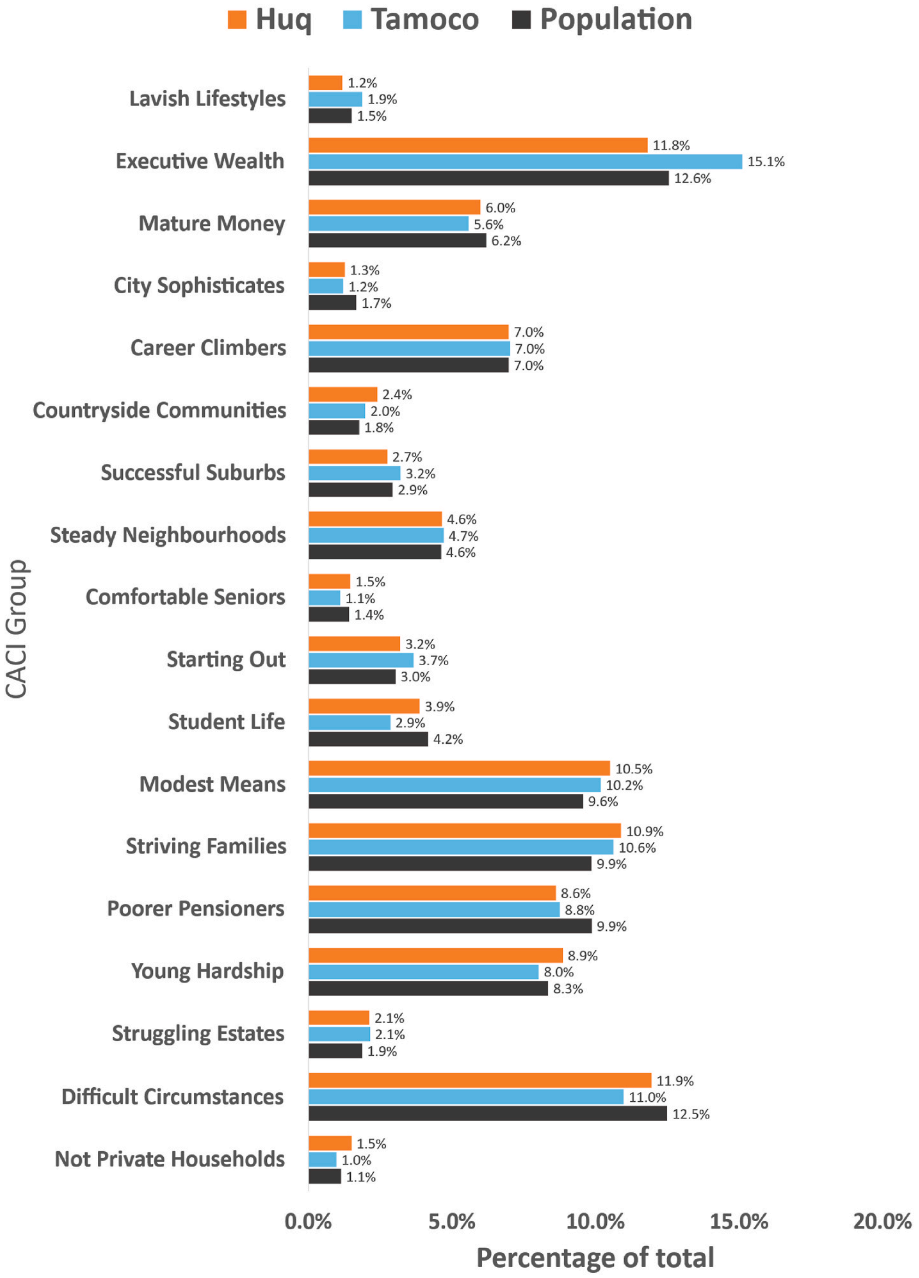


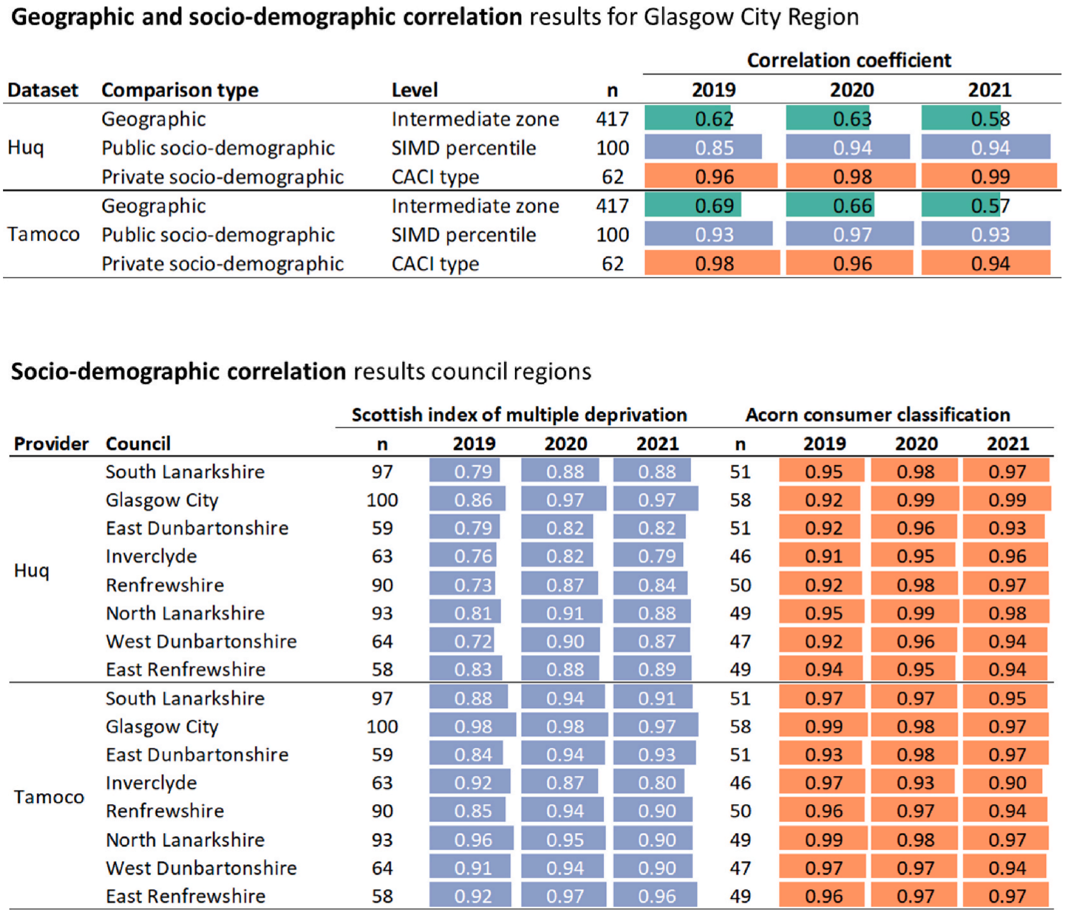
Fig. 4. Huq and tamoco by the 18 CACI groups in 2020.
Comparison is to the total population in 2020 (adult population data is not available at the level used for CACI data). Labels are percentage of the total population for each data source.

but this increases disclosure risks. The best approach therefore appears to be exploration of spatial coverage on the basis of estimated home locations with comparison to 'ground truth' sources such as official population statistics (Berke et al., 2022; Huang et al., 2022; Wang et al., 2019) as well as area-level measures of socio-demographic status (Bernabeu-Bautista et al., 2021; Huang et al., 2020).

Making judgements on the basis of area-level assessments does risk an ecological fallacy as we cannot claim to know the individual user characteristics from the area where they live. It is possible, for example, that we obtain good coverage of a 'poor' neighbourhood by capturing (atypical) 'richer' residents living within these locations. However, demonstrating even coverage across the spectrum of socio-demographic areas greatly reduces the potential for such an error. To obtain equal coverage in richer and poorer neighbourhoods without also having equal coverage of richer and poorer individuals, we would have to posit a mechanism whereby the richer residents in poorer neighbourhoods were more likely to be captured than richer residents in other kinds of place. Such a mechanism is complex and unwarranted so, following the principle of "Occam's razor" and taking the simpler explanation, we can reasonably conclude that good geographic coverage is a strong indication of good population coverage.

This study is the first to incorporate high resolution residential land use data into the process of individual home location detection for MPA data. Unlike more traditional forms of MP data such as call detail records, which are usually provided at a more aggregated level, or MP data from social media which are more fragmented, the high accuracy of MPA data makes the incorporation of building level residential data possible. Here we find that estimating home area by restricting data to that within buildings with a residential use has a positive impact on the socio-demographic fit of the data. By excluding locations related to

Table 3
Correlation between MPA users and population across the Glasgow city-region and within councils by geography and socio-demographic status.



Geographic and socio-demographic correlation results for Glasgow City Region

Dataset	Comparison type	Level	n	Correlation coefficient 2019	2020	2021
Huq	Geographic	Intermediate zone	417	0.62	0.63	0.58
	Public socio-demographic	SIMD percentile	100	0.85	0.94	0.94
	Private socio-demographic	CACI type	62	0.96	0.98	0.99
Tamoco	Geographic	Intermediate zone	417	0.69	0.66	0.57
	Public socio-demographic	SIMD percentile	100	0.93	0.97	0.93
	Private socio-demographic	CACI type	62	0.98	0.96	0.94

Socio-demographic correlation results council regions

		Scottish index of multiple deprivation				Acorn consumer classification			
Provider	Council	n	2019	2020	2021	n	2019	2020	2021
Huq	South Lanarkshire	97	0.79	0.88	0.88	51	0.95	0.98	0.97
	Glasgow City	100	0.86	0.97	0.97	58	0.92	0.99	0.99
	East Dunbartonshire	59	0.79	0.82	0.82	51	0.92	0.96	0.93
	Inverclyde	63	0.76	0.82	0.79	46	0.91	0.95	0.96
	Renfrewshire	90	0.73	0.87	0.84	50	0.92	0.98	0.97
	North Lanarkshire	93	0.81	0.91	0.88	49	0.95	0.99	0.98
	West Dunbartonshire	64	0.72	0.90	0.87	47	0.92	0.96	0.94
	East Renfrewshire	58	0.83	0.88	0.89	49	0.94	0.95	0.94
Tamoco	South Lanarkshire	97	0.88	0.94	0.91	51	0.97	0.97	0.95
	Glasgow City	100	0.98	0.98	0.97	58	0.99	0.98	0.97
	East Dunbartonshire	59	0.84	0.94	0.93	51	0.93	0.98	0.97
	Inverclyde	63	0.92	0.87	0.80	46	0.97	0.93	0.90
	Renfrewshire	90	0.85	0.94	0.90	50	0.96	0.97	0.94
	North Lanarkshire	93	0.96	0.95	0.90	49	0.99	0.98	0.97
	West Dunbartonshire	64	0.91	0.94	0.90	47	0.97	0.97	0.94
	East Renfrewshire	58	0.92	0.97	0.96	49	0.96	0.97	0.97

Results show Pearson's correlation coefficients, which are all significant at the 0.01 level. Results are based on method 1 (see methods). Geographic results are based on Intermediate Zones. SIMD results are based on Datazones grouped into percentiles. CACI Acorn results are based on postcodes grouped into 'types'. Geographic and SIMD comparison are to the adult population in 2020 while CACI comparison is to the total population in 2020. Tamoco results are the correlation with the mean estimated monthly users in a given year while Huq are based on the total users for a given year. Not all CACI types and SIMD percentiles are present in each council region, hence variations in the number of spatial units (n).

work, travel and leisure, our approach is more selective in estimating the home area over a standard technique. While this improvement comes with the minor caveat that fewer data points are attributed with a home area, the better socio-demographic fit offers an improved foundation from which to utilise the enriched data for analysis.

We demonstrate a strong representative from two samples but only for one city-region and one period of three years. With these kinds of data, representativeness will always need to be assessed on an on-going basis and our methodology is intended to help towards this end. Nevertheless, while we do not assume that the results in our case study area necessarily hold in other areas, there is also no reason to suggest that the Glasgow City Region, with its diverse socio-demographic composition, should return results which are exceptional. Therefore, the positive findings in relation to representativeness in this analysis should act as an encouragement for further research in the UK and potentially further afield. In this context, Huq currently offers data across the UK and Tamoco across the UK and the United States of America. Companies with a similar offering to the data assessed here are available internationally.

With survey response rates declining in recent years, new forms of MPA data may offer increasing potential as a complementary or alternative data source. In our case, the MPA data from two independent providers offers better coverage of the adult population than the unweighted data from the 'gold standard' household survey while also providing much greater sample size and coverage over multiple days (rather than the single day in the survey). Granted, more sophisticated weights can be applied to survey data to reduce existing biases. Nevertheless, starting from a more unrepresentative sample, these survey weights must do more work to compensate for sample limitations so the scope for bias to remain is that much greater. We do not suggest that MPA data are always better or that they offer a complete replacement for survey data; they cannot replace the detail on individual characteristics and attitudes which can be obtained during a survey, for example. Nevertheless, we would emphasise that concerns about potential bias in population coverage may have been greatly over-stated.

Despite the potential, there are two major factors which may still limit wider use of novel MPA data. The first is ethics and the regulatory requirements designed to protect individual privacy. There are clear ethical risks to privacy from location-based data with the kinds of spatio-temporal detail outlined here and regulators, at least in the UK, have been scrutinising how intermediaries like Huq and Tamoco handle these (Wakefield, 2021). While researchers' data management approaches need to pay equally careful attention to these, as with the current study, there is clear evidence that social science research use of MPA data does

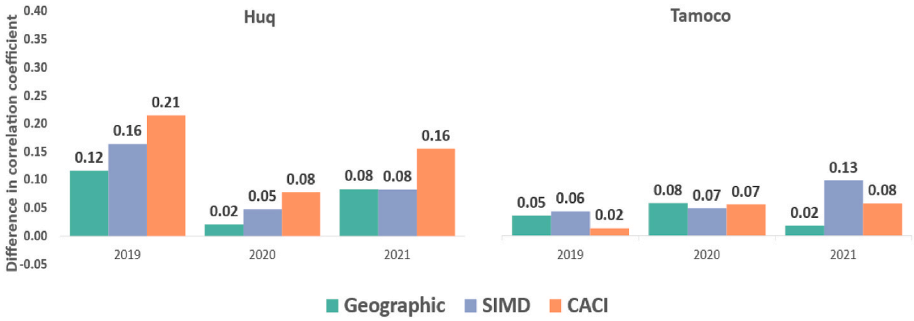


Fig. 5. Difference in correlations coefficient when using land use algorithm between number of MPA users and population across Glasgow City-region. Results based on Pearson's correlation coefficients for the relevant sample with 2020 populations. The geographic comparison is based on Intermediate Zones (417 regions). For SIMD, percentiles are used. For CACI, the 62 types are used. Tamoco results are based on the mean monthly users for a given year.

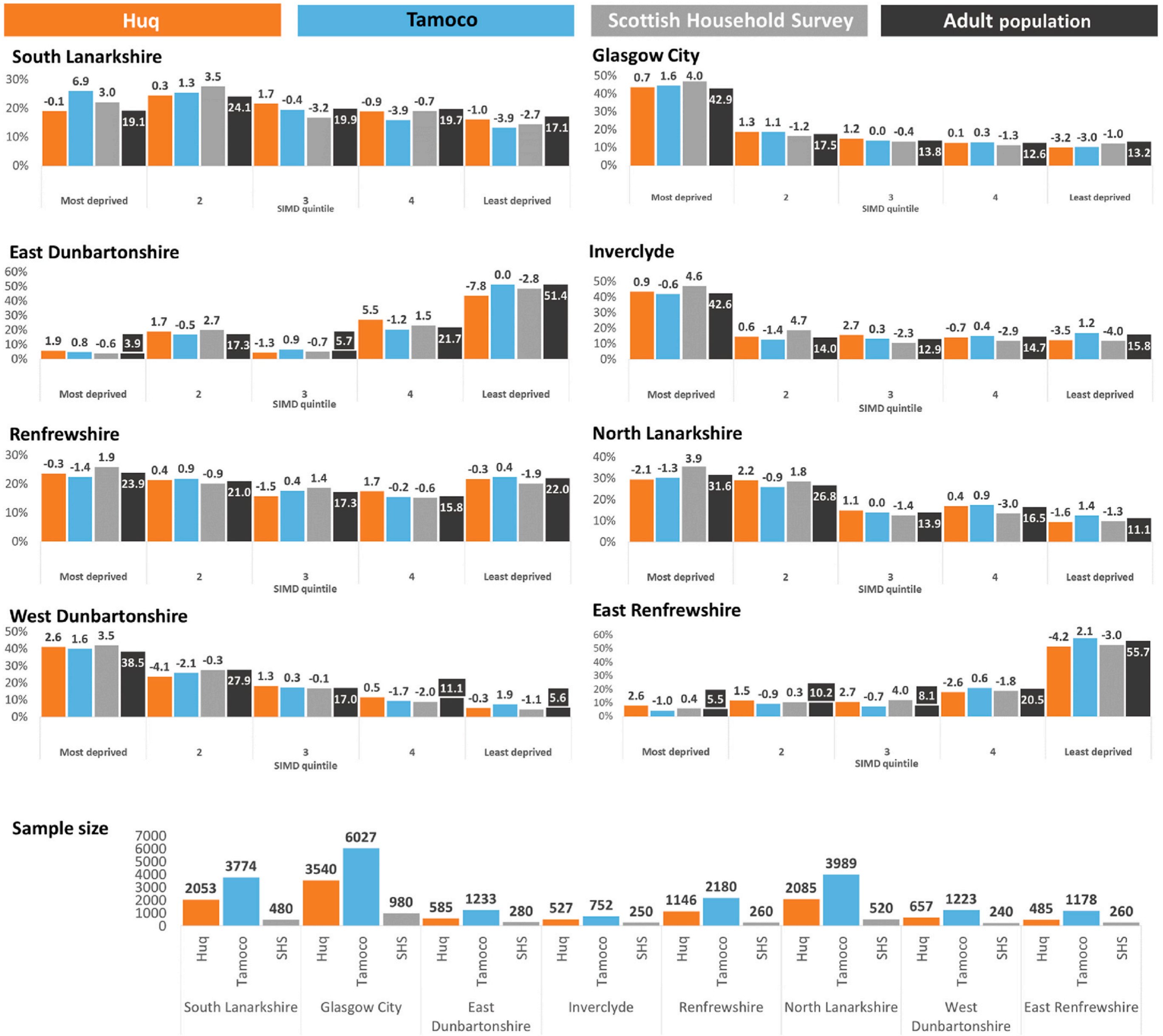


Fig. 6. Socio-demographic coverage of MPA data sample sizes compared with unweighted Scottish Household Survey samples across the eight councils of Glasgow City-region.
Population is the adult population in 2020. Data for the Scottish Household survey (SHS) are for 2019, the latest available. Results for both MPA datasets are based on our home location algorithm (Method 1) using 2020 data. Labels in the top 8 charts are percentage shares for the population figure and deviations from this for the others. Figures in the bottom panel show the MPA samples in 2020 and the number of SHS adult surveys in 2019 by council area. Tamoco figures are based on the mean monthly users in each case.

command widespread public support[5]. Researchers need to work with policy makers and regulators to ensure an appropriate balance is struck between individual rights to privacy and support for research with potential for public benefit. The second is cost, in terms of both license costs and data maintenance/processing costs. Although the details of licence agreements are usually confidential, it is commonly known that MP data in their various forms can be expensive to acquire (Lazer, 2006). This risks creating differentials in access, disadvantaging researchers from less well-funded institutions or countries. Intermediary data services like the UK's *Urban Big Data Centre* have a key role to play in relation to both, using central funding to secure licences which permit widespread use by other researchers at no further cost; and providing secure research facilities to reassure data owners that control will be maintained over the data and access limited to approved researchers and projects. Continued funding for such services will be essential to ensure a level playing field going forwards.

Here, we demonstrate sample representativeness of MPA data but future work should build on this to assess how well MPA data capture the movements of their sample (Ranjan et al., 2012; Zhao et al., 2016). With mobile data, location recording is not regular or scheduled, but sporadic and related to specific events such as sending or receiving a message (for call detail records) or using a specific app (for MPA data). These may lead them to capture some types of activity more than others. This may also vary between social groups so that, for example, while we may have a proportionate number of older people covered by the dataset, the ways in which they use apps may still lead to their movements being under- or over-represented. The precise mix of apps on which each company draws is commercially sensitive information and therefore not publicly available, but this clearly shapes outcomes as well as making them liable to change over time. Our findings on population representativeness should be seen, therefore, as the first stage in a larger process of evaluation.

Alongside the positive socio-demographic messages from this stage of our work, there are some limitations to note. First, while our approach should provide reassurance that we have good representativeness for social differences which vary over space such as income, age, or ethnicity, one area of representativeness this approach cannot cover is in relation to sex. Since men and women are, in general, very evenly spread geographically, it is not possible to pick up any under- or over-representation by sex when using an approach based on geographic variations. Given the widespread tendency for women to be under-represented in statistical data (Criado-Perez, 2020), this is an important priority for future work to address. Second, we can only identify home locations within the study area covered by the data used here. Residents from outside the city-region could be among the users for whom the algorithm fails to generate a home location (and hence omitted) or, if they spend even a few evenings in the city-region, they might be misattributed to a home location here. Incorporating residential land use, as well as the limiting of results to those with multiple evenings in residential space, should limit this error and should perform better than the previous approaches reliant solely on timing. In future work, we plan to explore this by extending the geographic scope of the analysis.

5. Conclusion

New sources of mobile phone application data offer unprecedented opportunities for applied scientific research. Given the novelty of these new forms of spatial data and their relative infancy as a data source, the future potential and direction of study are wide reaching. However, the ability to realise this potential is somewhat limited by uncertainties regarding sample representativeness. In this study, we have shown that mobile phone application data from two independent providers have a good fit to the population across public and private sources of socio-demographic data for a large city region which is home to over 1.8 million people. Furthermore, incorporating residential land use data into the process of home location detection for mobile phone data can improve its socio-demographic fit. These findings are important for future research as they present a technique which helps improves the fit of mobile phone application data while also offering an empirical foundation upon which to utilise these novel data sources for applied research.

Author statement

Michael Sinclair (MS), Saeed Maad (SM), Qunshan Zhao (QZ), Jinhyun Hong (JH), Andrea Ghermandi (AG) and Nick Bailey (NB). Conceptualization: MS, SM, QZ, JH, NB. Methodology: MS, QZ, JH, AG, NB. Software: MS, AG. Validation: MS. Formal analysis: MS. Visualization: MS. Writing - Original Draft: MS, QZ, JH, AG, NB. Writing - Review & Editing: MS, QZ, AG, NB.

Data statement

The mobile phone datasets used in this analysis can be accessed for research purposes by application to the Urban Big Data Centre, an Economic and Social Research Council funded research centre and national data service based at the University of Glasgow. The analysis is completed using a combination of PostgreSQL and R programming language (R Core Team, 2022). The code to process the data and estimate home location is openly available on GitHub (https://github.com/sinclairmichael/appliedgeography_representativeness.git).

Acknowledgments

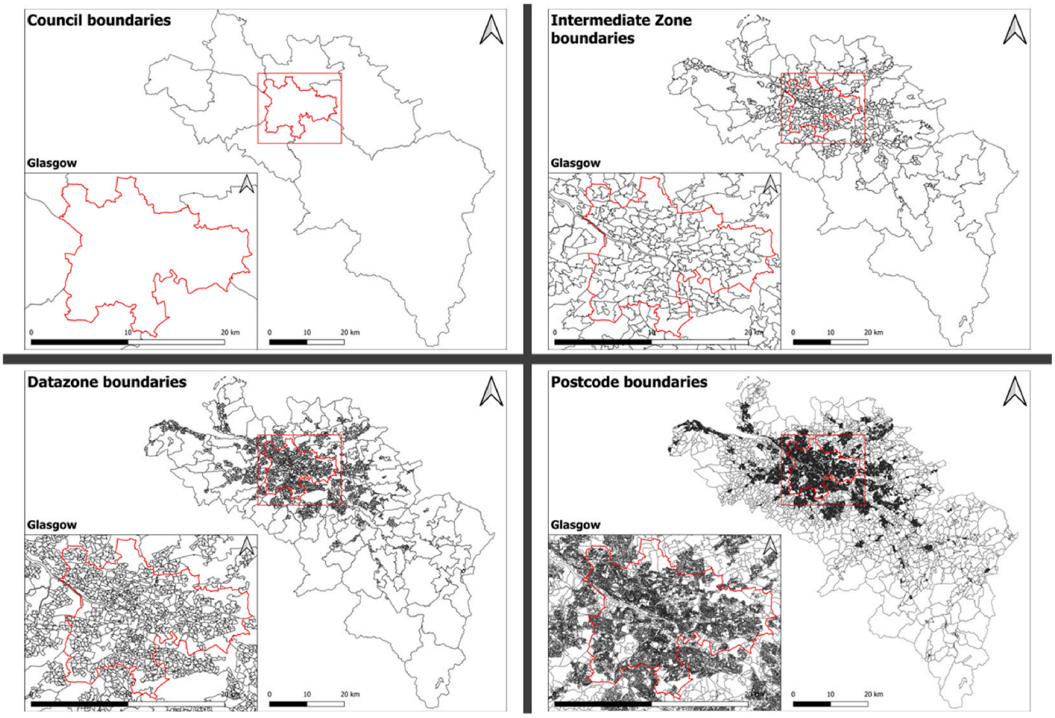
The work was made possible by ESRC's SDAI funding [ES/W012979/1] and ESRC's on-going support for the Urban Big Data Centre (UBDC) [ES/L011921/1 and ES/S007105/1]. For the use of SIMD data, Copyright Scottish Government, contains Ordnance Survey data © Crown copyright and database right (2022). CACI data, © 1979–2020 CACI Limited. This report shall be used solely for academic, personal and/or non-commercial purposes. UKBuildings data is Digital Map Data © The GeoInformation Group Limited (2022), created and maintained by Geomni, a Verisk company. The authors want to thank the anonymous reviewers for their insightful comments and suggestions on an earlier version of this manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.apgeog.2023.102997.

Appendix 1. Geographic boundaries used in the analysis and/or reporting of results

[5] https://www.gov.uk/government/publications/public-dialogue-on-location-data-ethics.



References

Berke, A., Doorley, R., Alonso, L., Arroyo, V., Pons, M., & Larson, K. (2022). Using mobile phone data to estimate dynamic population changes and improve the understanding of a pandemic: A case study in Andorra. *PLoS One, 17*, Article e0264860. https://doi.org/10.1371/journal.pone.0264860

Bernabeu-Bautista, Á., Serrano-Estrada, L., Perez-Sanchez, V. R., & Martí, P. (2021). The geography of social media data in urban areas: Representativeness and complementarity. *ISPRS International Journal of Geo-Information, 10*, 747. https://doi.org/10.3390/ijgi10110747

Bojic, I., Massaro, E., Belyi, A., Sobolevsky, S., & Ratti, C. (2015). Choosing the right home location definition method for the given dataset. In T.-Y. Liu, C. N. Scollon, & W. Zhu (Eds.), *Social informatics, lecture notes in computer science* (pp. 194–208). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-27433-1_14.

boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society, 15*(5), 662–679.

Brick, J. M., & Williams, D. (2013). Explaining rising nonresponse rates in cross-sectional surveys. *The Annals of the American Academy of Political and Social Science, 645*(1), 36–59.

Calabrese, F., Diao, M., Di Lorenzo, G., Ferreira, J., & Ratti, C. (2013). Understanding individual mobility patterns from urban sensing data: A mobile phone trace example. *Transportation Research Part C: Emerging Technologies, 26*, 301–313. https://doi.org/10.1016/j.trc.2012.09.009

Calafiore, A., Murage, N., Nasuto, A., & Rowe, F. (2021). Deriving spatio-temporal geographies of human mobility from GPS traces. *Spat. Data Sci. Symp*. https://doi.org/10.25436/E26K5F, 2021 Online.

Cameron, R. W. F., Brindley, P., Mears, M., et al. (2020). Where the wild things are! Do urban green spaces with greater avian biodiversity promote more positive emotions in humans? *Urban Ecosystems, 23*, 301–317. https://doi.org/10.1007/s11252-020-00929-z

Çolak, S., Alexander, L. P., Alvim, B. G., Mehndiratta, S. R., & González, M. C. (2015). Analyzing cell phone location data for urban travel: Current methods, limitations, and opportunities. *Transp. Res. Rec. J. Transp. Res. Board, 2526*, 126–135. https://doi.org/10.3141/2526-14

Criado-Perez, C. (2020). *Invisible women: Exposing data bias in a world designed for men*. London: Vintage.

Gao, S., Rao, J., Kang, Y., Liang, Y., Kruse, J., Dopfer, D., Sethi, A. K., Mandujano Reyes, J. F., Yandell, B. S., & Patz, J. A. (2020). Association of mobile phone location data indications of travel and stay-at-home mandates with COVID-19 infection rates in the US. *JAMA Network Open, 3*, Article e2020485. https://doi.org/10.1001/jamanetworkopen.2020.20485

Grantz, K. H., Meredith, H. R., Cummings, D. A. T., Metcalf, C. J. E., Grenfell, B. T., Giles, J. R., Mehta, S., Solomon, S., Labrique, A., Kishore, N., Buckee, C. O., & Wesolowski, A. (2020). The use of mobile phone data to inform analysis of COVID-19 pandemic epidemiology. *Nature Communications, 11*, 4961. https://doi.org/10.1038/s41467-020-18190-5

Guo, S., Song, C., Pei, T., Liu, Y., Ma, T., Du, Y., Chen, J., Fan, Z., Tang, X., Peng, Y., & Wang, Y. (2019). Accessibility to urban parks for elderly residents: Perspectives from mobile phone data. *Landscape and Urban Planning, 191*, Article 103642. https://doi.org/10.1016/j.landurbplan.2019.103642

Heo, S., Lim, C. C., & Bell, M. L. (2020). Relationships between Local Green Space and Human Mobility Patterns during COVID-19 for Maryland and California, USA. *Sustainability, 12*(22), 9401. https://doi.org/10.3390/su12229401

Huang, X., Li, Z., Lu, J., Wang, S., Wei, H., & Chen, B. (2020). Time-series clustering for home dwell time during COVID-19: What can we learn from it? *ISPRS International Journal of Geo-Information, 9*, 675. https://doi.org/10.3390/ijgi9110675

Huang, X., Lu, J., Gao, S., Wang, S., Liu, Z., & Wei, H. (2022). Staying at home is a privilege: Evidence from fine-grained mobile phone location data in the United States during the COVID-19 pandemic. *Annals of the Association of American Geographers, 112*, 286–305. https://doi.org/10.1080/24694452.2021.1904819

Kang, Y., Gao, S., Liang, Y., Li, M., Rao, J., & Kruse, J. (2020). Multiscale dynamic human mobility flow dataset in the U.S. during the COVID-19 epidemic. *Scientific Data, 7*, 390. https://doi.org/10.1038/s41597-020-00734-5

Kishore, N., Taylor, A. R., Jacob, P. E., Vembar, N., Cohen, T., Buckee, C. O., & Menzies, N. A. (2022). Evaluating the reliability of mobility metrics from aggregated mobile phone data as proxies for SARS-CoV-2 transmission in the USA: A population-based study. *Lancet Digit. Health, 4*, e27–e36. https://doi.org/10.1016/S2589-7500(21)00214-4

Lazer, D. (2006). Global and domestic governance: Modes of interdependence in regulatory policymaking. *European Law Journal, 12*, 455–468. https://doi.org/10.1111/j.1468-0386.2006.00327.x

Lee, K.-S., Eom, J. K., Lee, J., & Ko, S. (2021). Analysis of the activity and travel patterns of the elderly using mobile phone-based hourly locational trajectory data: Case study of gangnam, korea. *Sustainability, 13*, 3025. https://doi.org/10.3390/su13063025

Mao, H., Shuai, X., Ahn, Y.-Y., & Bollen, J. (2015). Quantifying socio-economic indicators in developing countries from mobile phone communication data: Applications to côte d'Ivoire. *EPJ Data Sci, 4*, 15. https://doi.org/10.1140/epjds/s13688-015-0053-1

Marsh, C. (1982). *The survey method: contribution of surveys to sociological explanation*. London: Allen & Unwin.

Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt.

Mears, M., Brindley, P., Barrows, P., Richardson, M., & Maheswaran, R. (2021). Mapping urban greenspace use from mobile phone GPS data. *PLoS One, 16*, Article e0248622. https://doi.org/10.1371/journal.pone.0248622

Meyer, B. D., Mok, W. K., & Sullivan, J. X. (2015). Household surveys in crisis. *Journal of Economic Perspectives, 29*(4), 199–226.

National Records of Scotland. (n.d.). National Records of Scotland | Preserving the past, Recording the present, Informing the future. https://www.nrscotland.gov.uk/.

Pappalardo, L., Ferres, L., Sacasa, M., Cattuto, C., & Bravo, L. (2021). Evaluation of home detection algorithms on mobile phone data using individual-level ground truth. *EPJ Data Sci, 10*, 29. https://doi.org/10.1140/epjds/s13688-021-00284-9

Phithakkitnukoon, S., Smoreda, Z., & Olivier, P. (2012). Socio-geography of human mobility: A study using longitudinal mobile phone data. *PLoS One, 7*, Article e39253. https://doi.org/10.1371/journal.pone.0039253

Ranjan, G., Zang, H., Zhang, Z.-L., & Bolot, J. (2012). Are call detail records biased for sampling human mobility? *ACM SIGMOBILE - Mobile Computing and Communications Review, 16*, 33–44. https://doi.org/10.1145/2412096.2412101

R Core Team. (2022). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Ren, X., & Guan, C. (2022). Evaluating geographic and social inequity of urban parks in Shanghai through mobile phone-derived human activities. *Urban Forestry & Urban Greening, 76*, 127709.

Savage, M., & Burrows, R. (2007). The coming crisis of empirical sociology. *Sociology, 41* (5), 885–899.

Scottish Government, Ipsos MORI. (2021). SHSScottish household survey. *1999-Scottish Household Survey*. https://doi.org/10.5255/UKDA-SN-8775-1, 2019.

Sinclair, M., Mayer, M., Woltering, M., & Ghermandi, A. (2020). Using social media to estimate visitor provenance and patterns of recreation in Germany's national parks. *Journal of Environmental Management, 263*, Article 110418. https://doi.org/10.1016/j.jenvman.2020.110418

Sinclair, M., Zhao, Q., Bailey, N., Maadi, S., & Hong, J. (2021). Understanding the use of greenspace before and during the COVID-19 pandemic by using mobile phone app data. GIScience 2021 Short Pap. In *Proc. 11th int. Conf. Geogr. Inf. Sci. Sept. 27-30 2021*. https://doi.org/10.25436/E2D59P. Poznań, Poland (Online).

Vanhoof, M., Reis, F., Ploetz, T., & Smoreda, Z. (2018). Assessing the quality of home detection from mobile phone data for official statistics. *J. Off. Stat., 34*, 935–960. https://doi.org/10.2478/jos-2018-0046

Wakefield, B. J. (2021, October 29). Location data collection firm admits privacy breach. *BBC News*. https://www.bbc.co.uk/news/technology-59063766.

Wang, F., & Chen, C. (2018). On data processing required to derive mobility patterns from passively-generated mobile phone data. *Transportation Research Part C: Emerging Technologies, 87*, 58–74. https://doi.org/10.1016/j.trc.2017.12.003

Wang, Y., Li, J., Zhao, X., Feng, G., & Luo, X. (2020). Using mobile phone data for emergency management: A systematic literature Review. *Information Systems Frontiers, 22*, 1539–1559. https://doi.org/10.1007/s10796-020-10057-w

Wang, F., Wang, J., Cao, J., Chen, C., Ban, X., & Jeff). (2019). Extracting trips from multi-sourced data for mobility pattern analysis: An app-based data example. *Transportation Research Part C: Emerging Technologies, 105*, 183–202. https://doi.org/10.1016/j.trc.2019.05.028

Yabe, T., Jones, N. K. W., Rao, P. S. C., Gonzalez, M. C., & Ukkusuri, S. V. (2022). Mobile phone location data for disasters: A review from natural hazards and epidemics. *Computers, Environment and Urban Systems, 94*, 101777.

Yabe, T., Tsubouchi, K., Fujiwara, N., Sekimoto, Y., & Ukkusuri, S. V. (2020). Understanding post-disaster population recovery patterns. *Journal of The Royal Society Interface, 17*, Article 20190532. https://doi.org/10.1098/rsif.2019.0532

Zhao, Z., Shaw, S.-L., Xu, Y., Lu, F., Chen, J., & Yin, L. (2016). Understanding the bias of call detail records in human mobility research. *International Journal of Geographical Information Science, 30*, 1738–1762. https://doi.org/10.1080/13658816.2015.1137298