The impact of privacy protection measures on the utility of crowdsourced cycling data

Varun Raturi[a], Jihnyun Hong[a], David Philip McArthur[a], Mark Livingston[a]
[a] *Department of Urban Studies, The University of Glasgow, Glasgow, UK*

Abstract

The use of new forms of data in the transport research domain is rapidly gaining popularity. However, these data come with specific challenges and one of the major concerns is maintaining the privacy of data subjects. One widely used approach to anonymise data is to apply binning. Recently, data from activity-tracking applications like Strava has been utilised to study active travel. Due to privacy concerns, Strava has started providing data in a discretised format from July 2018. In this study, we aim to analyse the impact of the binning criteria on the utility of the crowdsourced data by using Strava Metro data from 2013 to 2016 for the city of Glasgow. We applied the Strava binning criteria on the original, disaggregate dataset at three different temporal aggregations (i.e., Hourly, Daily and Monthly) and conducted different analyses to examine the impacts. First, we compared manual cycling counts with the original and binned cycling counts from Strava. Second, net-errors were calculated by comparing original and binned cycling counts from Strava data. Third, we estimated spatial autocorrelation statistics based on original and binned Strava counts and investigated the extent to which research outcomes change because of the binning approach. Our results confirmed significant information loss. Worryingly, we also show that conclusions reached by previous studies could have been reversed if the new specification of the data had been used. We outline here what precautions researchers and planners should take when working with the binned data.

Keywords: *Cycling, Privacy, Infrastructure, Crowdsourced data, Strava, Spatial Autocorrelation*

1. Introduction

Cycling has become a key research topic due to the significant benefits it can generate, such as reducing car-dependency, air pollution and congestion (Chapman et al., 2018). Some studies have indicated that a travel behaviour change in favour of active travel can be used as a climate change mitigation strategy (Rissel, 2009; Woodcock et al., 2009) and can have a positive impact on the economy (Cavill, 2008; Mindell, 2015). Although cycling to work increases the risk of injuries (Welsh et. al., 2020), there is a clear consensus on the net health benefits of travel by bicycle (Celis-Morales et al., 2017; De Hartog et al., 2010; Mueller et al., 2015; Oja et al., 2011). These substantial benefits have caught the attention of policy makers and there is a push to increase the bicycle mode share.

To achieve an increase in cycling mode share, it is essential to understand current cycling patterns and develop valid appraisal and evaluation tools. Doing this requires accurate and reliable data. However, a lack of suitable data prevents us achieving a general understanding of the travel patterns and behaviour of cyclists. Traditionally cycling data is collected through travel surveys or cycling counts. Household surveys tend to be expensive and in order to capture a representative sample of cyclists, many households need to be surveyed (due to the low mode share of cycling). While manual cycling count data are a good indicator of the overall volume, they are temporally and spatially sparse. Moreover, malfunctioning automatic counters can lead to biased estimates of cycling volumes

due to gaps in the data (El Esawey et al., 2015). These traditional data sources are also less likely to capture detailed route information of cyclists.

Due to the limitations of traditional sources of cycling data, there has been a surge in data-driven research on cycling using new forms of data (Romanillos et al., 2016). Researchers have used sources such as volunteered geographic information (VGI) to analyse cyclist route information (Harvey & Krizek, 2007; Keler & Mazimpaka, 2016; Prandi et al., 2014); GPS data to extract walking and cycling trips (Eisenman et al., 2009; Stopher et al., 2008); bike share data to study cyclists' journeys and their relation with public transport (Froehlich et al., 2009; Padgham, 2012); and activity-tracking app data to examine cycling patterns (Hong et al., 2020; Jestico et al., 2016; Norman et al., 2019).

Although much of the discussion in the new forms of data context highlights the enormous potential of the data and its transformational value for research, it also raises some concerns and challenges. Aside from representativeness, one of the major concerns is privacy (Nunan & Di Domenico, 2017; Rubinstein, 2013; Tene & Polonetsky, 2012). Some studies in the field of transport, especially emerging technologies, have also raised similar concerns (Kitchin, 2016; Lederman, Taylor, & Garrett, 2016). One privacy protection approach used by crowdsourced data providers is to apply some form of aggregation to the raw data collected from the data subjects. However, there exists a trade-off between privacy and data utility and hence efforts to protect data subjects' identities can lead to loss of information, potentially influencing research outcomes (Lin, Hewett, & Altman, 2002; J. R. Roy, Batten, & Lesse, 1982).

The increasing popularity of activity-based social networks provides an opportunity to collect crowdsourced data from cyclists and pedestrians, and thus gain insight into their travel behaviour. Among the various activity-tracking applications, Strava is one of the most popular, having collected users' data since 2009. Strava Metro (the suite of data products derived from the Strava app) made some of this data available to researchers and planners. One of the data products provided minute-by-minute activity counts for each road link. However, due to mounting privacy concerns, Strava Metro has started providing the activity counts on each road link in a binned format (from July 2018) by aggregating the data in five-count buckets and lowering the temporal resolution of the data. The Strava binning criteria rounds up the activity counts on every link to the nearest multiple of 5. For example, 4 becomes 5 and 8 becomes 10. Any count less than or equal to 3 is rounded down to zero. Strava does not provide any reasoning behind the adoption of this binning strategy. Furthermore, based on the fact that in many places there are only small numbers of people who cycle, and that only a portion of cyclists use the app, their binning strategy could result in substantial information loss. Conventional binning is less effective if the dataset contains a lot of smaller values as the level of information lost will be high (Hundepool et. al, 2010). This is important especially for transport planners because several local authorities have used Strava data to make cycling plans (e.g., Strava signed a data sharing contract with Transport for London in 2016 and renewed the contract in 2019). This paper aims to analyse some of the potential impacts of these binning strategies on the utility of the crowdsourced data and robustness of analyses using these data. Specifically, we focus on four research questions (RQ) analysing the missing activities and the reduction in validity of the binned dataset. We also investigate the impact on the research outcomes due to the binning strategy adopted.

We used Strava Metro data covering the period 2013 to 2016 to examine the extent to which impacts of binning may vary over time. Although a significant number of studies highlight the relation between data aggregation and

loss of information, to the best of our knowledge, the impact of these binning criteria on the quality of crowdsourced data, especially in the context of cycling research, has not yet been examined. A number of recent literatures have used crowdsourced data for cycling research; especially in the field of transport geography and transport policy; therefore, realising the limitations of the data used is critical, given that this kind of data is being used to inform policy. Furthermore, utilising four years of crowdsourced data makes this study a significant contribution to the literature.

The paper is structured as follows. Section 2 reviews previous studies of cycling which have utilised Strava Metro data. These studies utilised the original version of the data and would potentially be affected with the new, binned data. Section 3 describes the data and gives details of the methodology adopted. Each research question is addressed using different methods. Section 4 presents the results which are discussed in Section 5.

2. Literature Review

Strava is an online/smartphone application for recording and monitoring physical activity and allows the user to record and upload their activity based on their device's location data. Recently, there has been a surge in studies investigating a diverse range of topics using Strava data.

Several authors have compared cycling volumes from Strava with manual cycling counts to establish the usability of the Strava data for cycling behaviour studies and consequently validating their use in spatial analyses (Perkins and Blake, 2016; Griffin & Jiao, 2015; Livingston et al., 2020; Boss et al., 2018). Jestico et al. (2016) computed the linear association of the Strava data with the cordon count data at 18 locations in Victoria, Canada. In the study, a peak period aggregation was used as it had higher R^2 values compared to hourly aggregations. Using Strava counts as one of the explanatory variables along with variables like speed limit, slope and on-street parking, this study predicted cycling volumes (split into categories) for the city of Victoria using a generalised linear model and found that users of crowdsourced apps chose similar routes to commuter cyclists in an urban environment. They also concluded that incorporating spatial and temporal detail along with crowdsourced data will provide better insight into cyclists' behaviour. Haworth (2016) used ordinary least squares models to estimate London census cycle flows using Strava activity counts and concluded it was a good predictor with an R^2 of around 0.62.

Strava data have been used by researchers to estimate the bicycle ridership at the city level, and can be extremely valuable to analyse cyclist travel behaviour due to its rich temporal and spatial content. Griffin & Jiao (2015) used Strava data to analyse the travel behaviour of cyclists who have the pursuit of fitness as one of their motivations to cycle. Using ordinary least squares and geographically weighted regression, bicycle kilometres travelled were predicted and the model showed that cyclists tend to use streets that are in diverse and populated regions and have steeper slopes. Similarly, Hochmair, Bardin, & Ahmouda (2019) explored the differences between commute and non-commute bicycle kilometres travelled (BKT) using a five-month Strava dataset for the Miami-Dade County area. They also investigated the difference between weekend and weekday BKT using the dataset aggregated for weekdays and weekends. In this study, the ridership was defined as the dependent variable and was a function of road network, built environment and sociodemographic variables. Proulx & Pozdnukhov (2017) fused the Strava data with other data sources such as count data, bike-share data, and outcomes from two travel demand models using a Geographically Weighted Data Fusion (GWDF) technique based on Geographically Weighted Regression.

In this study, Strava data primarily served the purpose of representing recreational travel. The final fused dataset was then used to estimate link-level bicycle flows on a network.

Several studies have examined the travel safety and pollution exposure for pedestrians and cyclists using Strava data (Chen et al., 2017; Lee & Sener, 2019; Saad, Abdel-Aty, Lee, & Cai, 2019; Sanders et al., 2017; Wang et al., 2018). Saha et al. (2018) used Strava data for the year 2014 for Florida to compute the bicycle trip miles and bicycle trip intensity for each census block. These measures were then incorporated as covariates in two conditional autoregressive models to predict bicycle crashes. Sun & Mobasheri, (2017) explored the potential of using Strava data to determine the air pollution exposure of the cyclists segregated by their trip purposes. The study used data from Glasgow and compared commuting trips with non-commuting trips and concluded that recreational trips are less prone to air pollution exposure as they generally happen at the outskirts of the city and highlighted the need for policymakers to improve cycle infrastructure on the periphery of the city.

Strava data has been used in various spatial analyses of the spatial patterns of cyclists (Boss et al., 2018; Campbell et al., 2019; Griffin & Jiao, 2015; Orellana & Guerrero, 2019). To understand how conventional and crowdsourced data represent actual cycle ridership, Conrow et al., (2018) used local indicator of spatial association (LISA) to determine the clusters and outliers based on differences in ridership obtained from Strava data and conventional count data. To compare the two datasets, the Strava link flows segregated by direction were converted to counts at the location of the manual counts. These clusters and outliers were then correlated with variables like population density, road density, land-use data etc. to provide some comparison between Strava users and all riders. They found that areas with low density, greater social disadvantage and lower ridership had strong spatial similarity. McArthur & Hong (2019) used the sample of commuters from the Strava dataset for the city of Glasgow to analyse commuting cyclist travel patterns. They used the Strava OD flows and applied an all or nothing traffic assignment model to determine the shortest link flows. This was then compared to the actual flows hence shedding light on the cyclists' commuting behaviour by concluding that cyclist might prefer routes with higher scenic value and good cycling infrastructure.

Strava data has been also used for infrastructure assessment analysis. Hong et al., (2020) assessed the impact of new cycling infrastructure in Glasgow on ridership by using fixed effects panel data regression models on Strava data for the years 2013-2016. The conclusion of the study was that majority of the projects lead to an increase in cycling volumes. Boss et al., (2018) calculated spatial autocorrelation measures on the change in Strava activity counts and used the clusters and outliers on links to determine whether the change in activities exhibit spatial dependence. By overlapping these results with the infrastructure improvements on those links, the study assessed the impact of these investments.

Although Strava data has been used for diverse range of objectives, the limitations of the Strava dataset are commonly accepted. One of the major issues with the dataset is that the sample of cyclists is biased (McArthur & Hong, 2019; Sanders et al., 2017), and some studies have even attempted to correct these biases (A. Roy et al., 2019; Saad et al., 2019).

Strava's new approach to privacy protection rounds down any count less than or equal to 3 to zero. Some literature suggests that for every 51 cyclists observed on a link, only one will use Strava (Jetsico et al. 2016). Consequently, a link with over 150 cyclists may be reported as having no cycling activity in the Strava data. This can influence

research outcomes significantly due to the information loss. However, no empirical studies have examined to what extent this new approach (binning) could influence research outcomes. Because of the popularity of Strava data for research and local cycling planning, it is critical to understand the potential impacts of the change in the specification of the data.

3. Data and Methodology

We used the city of Glasgow as our case study. Glasgow is Scotland's largest city with a population just under 600,000. There has been a rise in vehicle-km travelled due to increasing vehicle ownership. Cycling volumes have also increased, but the increase was not enough to meet the vision of having 10% of everyday journeys by bicycle by 2020. In 2015-2016, within Glasgow, only 3.9% of people commuted to work or study by bicycle compared to 6.7% in Edinburgh (Annual Cycling Monitoring Report, Cycling Scotland, 2018). Hence, significant efforts and investments are currently being made to increase this share by making cycling safer and more accessible (Glasgow City Council, 2016).

3.1. Data sets

3.1.1. Strava Data

We utilised data provided by Strava Metro which is derived from consenting users of the Strava activity-tracking app from 2013 to 2016 for the city of Glasgow. Strava provides an OpenStreetMap shapefile of the city where activity counts are provided for each link, thus, we used minute-by-minute cycling count data for each link. The dataset also contains aggregated demographic information (Table 1). It is worth noting that there are relatively more male cyclists in the Strava Metro data (87%) compared to the National Travel Survey[1] (71%).

Table 1: Summary of the Strava data for Glasgow (2016)

Age	Male	Female
< 20	847	207
25 – 34	2332	544
35 – 44	3316	456
45 – 54	2675	337
55 – 64	720	78
65 – 74	110	5
75 – 84	8	0
85 – 94	2	0
Unknown	3075	580
Total	13085	2207

Table 2 shows the total cycling counts over 4 years. The total counts increased continuously. The activity counts (one cyclist can have multiple activities) are calculated by summing activities over each link in the network. A more robust way to compute total activities can be by utilising the Origin-Destination data as the activities number might change if the representation of the network changes (i.e., if a road represented by one link is subsequently represented by two links, then the activity count may double). However, the aim of the research is to determine how these counts are affected by binning rather than to produce a series which is consistent over time.

[1] https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/577825/mode-use-2015-a-view-into-a-travel-week.pdf

Table 2: Total activity counts for cyclists in Glasgow

Year	Total activity count
2013	14,567990
2014	22,819663
2015	31,561698
2016	35,003961

3.1.2. Manual cycle counts (from a cordon count)

To measure the suitability of Strava Metro data for capturing cyclist activity, its relationship with ground truth data needs to be established. Ground truth data of cyclist activities are captured by conducting cordon count surveys. Glasgow City Council has conducted a two-day cordon count every September since 2007 (Livingston et. al. 2020). For this study we used cordon counts from the year 2014, carried out from 6 am to 8 pm at 38 points around the city centre. The cordon count locations are shown in Figure 1.

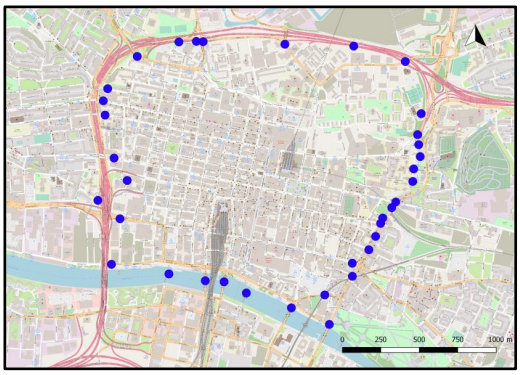

Figure 1: Cordon count locations; Source: Glasgow city council

3.2. Methodology

Four research questions are investigated in this paper and different approaches have been used for each. As mentioned, the Strava Metro data before July 2018 include disaggregated minute-by-minute cycling counts.

In this study, we applied the aggregation rules to the disaggregated dataset previously supplied by Strava Metro at three different temporal aggregation levels (hourly, daily and monthly) for the years 2013 to 2016. Hence, minute-by-minute activity counts on each link were aggregated to hourly, daily and monthly counts. Then the Strava binning criteria were applied at all these temporal aggregations. We refer to the resulting datasets as *Original Strava data* and *Binned Strava data* for all the temporal aggregations for four years.

RQ1) To what extent does the binning approach reduce the validity of the data?

To answer this research question, we compared total activity counts from Strava Metro data (both *original* and *binned*) with cordon count data by utilising two different correlation coefficients. We calculated hourly (from 6am to 8pm) and daily cycling counts from cordon count data and corresponding Strava Metro data (both *original* and *binned* datasets). We calculated Spearman's rank and Pearson correlation coefficient to examine the extent to which the binning approach reduces the correspondence between Strava data and the ground truth.

***RQ2)* How many activities are missed in the counts due to the binning criteria?**

We computed the errors generated by rounding down counts of three or less (round-down errors) and by rounding up other counts to the nearest multiple of five (round-up errors). For this analysis, we used total activity counts aggregated by hour, day and month.

The summation of these errors implies the net-errors as these values deviate from the value in the *original* dataset. We examined the percentage of round-up and round-down errors across the temporal aggregations in the binned dataset.

***RQ3)* How do the changes in activity counts differ between the original and the binned dataset?**

To answer this question, we examined the changes in cycling volumes from the *original* and *binned* Strava datasets across years. As previous studies used the changes in cycling volumes for infrastructure impact analysis (Boss et al., 2018; Hong et al., 2019), analysing how these changes in activity counts differ between both datasets becomes crucial. The objective of this step is to determine whether the *binned* data can capture similar changes in ridership on links compared to the *original* data. For this, we used the activity counts from 2013 and 2014. First, using the original dataset, activity counts on every link from the 2013 data were subtracted from the activity counts on the corresponding links from the 2014 data. Similarly, changes in activity counts for every link between the given years were calculated with the *binned* data. Second, we compared whether results from the *original* and *binned* datasets follow the same trend (i.e., increase or decrease).

***RQ4)* To what extent do the research outcomes of cycling infrastructure impact analyses change due to the binning strategy?**

To address this question, we adopted the methodology proposed by Boss et al., (2018) to examine impacts of cycling infrastrucutre. We applied the methdology to Glasgow using both the *original* and *binned* data, and compared the results. This involves testing for spatial dependence in cycle counts. Spatial autocorrelation measures capture both attribute and locational similarity. Thus, we calculated spatial autocorrelation on the change in bicycle activity counts to determine whether there are any spatial patterns.

Specifically, we selected one of the new cycling investments in Glasgow (namely Cathkin 1, which was opened in October 2013) and calculated spatial autocorrelation measures to examine whether the changes in the total cycling volumes made by Strava users are significantly different from the random changes around the new cycling path. As a way to identify local clusters and outliers, concept of local indicator of spatial association (LISA) statistic was suggested by Anselin, (1995). The LISA for each observation gives an indication of the extent of significant spatial clustering of similar values around that observation. One such local indicator of spatial association is the Local Moran's I.

Local Moran's I is a common measure to determine statistically significant high clusters and outliers by computing high and low values relative to the mean. Local Moran's I is formulated as:

$$I_i = \frac{n(x_i - \bar{x}) \sum_j^n w_{ij}(x_j - \bar{x})}{\sum_i^n (x_i - \bar{x})^2}$$

where n is the number of regions, x_i is the attribute value at area i, x_j is the attribute value at the neighbouring area j and the spatial relationship between areas is determined by the weight matrix w_{ij}. $\bar{x}$ is the attribute mean value across all the areas.

Spatial weight matrix is an imposed structure to the data to determine the neighbours of each location/entity. We formulate our spatial weight matrix based on network contiguity. Here links are the basic entity of analysis and each link's neighbours are determined based on their connectivity. In this study, a binary adjacency matrix is used following Boss et al., (2018), where $w_{ij} = 1$ for links that are connected directly to each other, otherwise $w_{ij} = 0$. We generated the weight matrix for the Glasgow city network using Geoda software[2]. As mentioned earlier, to measure the impacts of cycling infrastructure, the changes in cycle activities on these links from 2013 to 2014 were used because the infrastructure was built in 2013. We applied Local Moran's I to the changes in activities between 2013 and 2014 to determine whether the results vary according to the *original* and the *binned* datasets, potentially leading to different research conclusions.

4. Results

In this section, we present the results for each research question derived from the methods and data outlined in Section 3. By answering these research questions, we are better able to understand the overall impact of the binning approach adopted in the *binned Strava data on cycling studies*.

4.1 RQ1) To what extent does the binning approach reduce the validity of data?

The Spearman's rank correlation coefficient and the Pearson correlation coefficient between cordon counts and the two Strava datasets are presented in Table 3. Given the ordinal nature of binned data, a Spearman's rank coefficient might be more suitable giving the strength and direction of the monotonic association between the variables whereas Pearson correlation coefficient indicates the linear relationship. Results show high positive correlations even for *original* hourly aggregated cycling counts. This implies that *original* Strava datasets have potential value for examining the spatial patterns of cycling behaviour. However, we do find that these correlation coefficients become smaller when using *binned* Strava data compared to the *original* dataset, indicating the reductions of the ability of the data to capture cycling flows, especially binned hourly aggregation dataset which shows the lowest correlation value of around 0.35 and 0.721 for the Spearman's rank and Pearson respectively. Figure 2 shows the scatterplots between cordon counts and Strava counts (original and binned datasets). We can see positive correlations but a substantial number of zero activities in the *binned* data, could result in lower association with the cordon counts. As Strava datasets have been used in mapping cycle ridership using regression models, a lower association with the cordon counts for *binned* datasets can reduce their applicability and accuracy for this area of research too. Since most roads in the city have a very low number of cycling activities (e.g., under 5 per hour), using *binned* data might introduce a significant bias in cycling research.

[2] Geoda is a free and open-source spatial analysis software. The currently released version of the software cannot read network shape files and hence at the time of writing the paper, this version of the software (capable of reading network shapefiles) has not been publicly released and was received by personal contact with the developers.

Table 3: Correlation coefficients

Variables (Cordon counts- Strava counts)	Spearman's Rank	Pearson
Original Hourly aggregation	0.5294	0.793
Binned Hourly aggregation	0.3589	0.721
Original Daily aggregation	0.7307	0.9
Binned Daily aggregation	0.7011	0.863

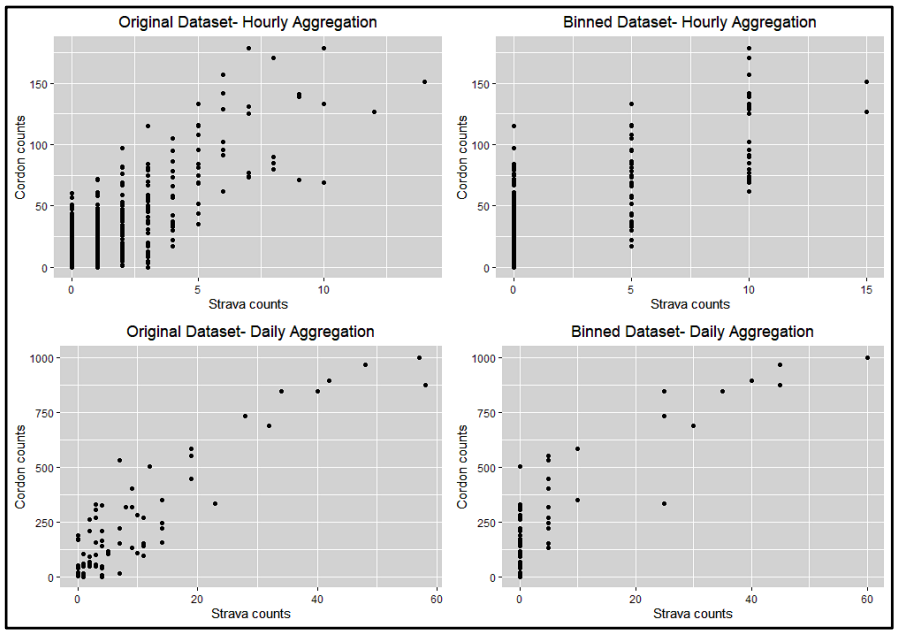

Figure 2: Scatterplot: Cordon Counts and Strava counts

4.2 RQ2) How many activities are missed in the counts due to the binning criteria?

As mentioned in the methodology, we applied the Strava binning criteria to the *original*, disaggregate data. We then calculate round-up and round-down errors to assess the ability of the binned data to measure the total level of cycling activity in Glasgow. The results are shown in Table 4.

Table 4: Net-error due to Strava binning criteria

Year	Aggregation	Activities		Errors					
		Original	Binned	Round-down (Abs)	Round-down (%)	Round-up (Abs)	Round-up (%)	Net (abs)	Net (%)
2013	Hourly	14,567990	1,981825	12,934377	88.79	348212	2.39	13,282589	91.18
	Daily	14,567990	11,491025	5,054648	34.70	1,977683	13.58	7,032331	48.27
	Monthly	14,567990	14,858215	211904	1.45	502129	3.45	714033	4.90
2014	Hourly	22,819663	4,313765	19,314154	84.64	808256	3.54	20,122410	88.18
	Daily	22,819663	20,252520	5,819411	25.50	3,252268	14.25	9,071679	39.75
	Monthly	22,819663	23,186450	221879	0.97	588666	2.58	810545	3.55
2015	Hourly	31,561698	8,222530	24,927277	78.98	1,588109	5.03	26,515386	84.01
	Daily	31,561698	29,189845	6,648418	21.06	4,276565	13.55	1,0924983	34.61
	Monthly	31,561698	32,008755	249743	0.79	696800	2.21	946543	3.00
2016	Hourly	35,003961	10,040395	26,877759	76.78	1,914193	5.47	28,791952	82.25
	Daily	35,003961	33,170065	6,509250	18.60	4,675354	13.36	11,184604	31.95
	Monthly	35,003961	35,459860	230312	0.66	686211	1.96	916523	2.62

For the 2013 data, the results show that the activity loss at the hourly aggregation level due to the round-down criteria is as high as 89 percent. This is due to the low number of activities in some parts of the city at the hourly level. Anywhere with three or fewer activities is rounded down to zero. The round up error is smaller, at around 2.4 percent. Thus, the net error at this temporal aggregation is approximately 91 percent. The round-down error reduces the ability to detect activity significantly for the daily aggregation however the round-up error increases

to 13.5 percent. The total net error reduces and is around 50 percent. For the monthly aggregation, both round-down and round-up errors reduce, and the net error is around 5 percent. Overall, monthly aggregation retains the maximum information compared to the original dataset due to the relatively large volume of cycling activity which takes place over the course of a month. The results imply that the *binned* Strava datasets are not appropriate to examine aggregated measure of activity especially at hourly or daily level.

Across different years, the net error for all the temporal aggregation reduces due to the increased number of Strava users in Glasgow. When the total number of cycling activities increased from around 14.5 million to 35 million, the net errors reduced to around 10 percent for the hourly aggregation. The net loss for daily and monthly aggregations also reduces, indicating that as number of Strava users increases, the data can retain more activity count information under this binning criteria. This suggests that the value of *binned Strava data* will be lower for cities with low bicycle mode share and a smaller number of Strava users compared to cities with high cycling mode shares and lots of Strava users.

4.3 RQ3) How do the changes in activity counts differ between the original and the binned data?

As shown in Table 4, hourly and daily aggregations show high net errors. Therefore, we examined the changes in activities at the hourly and daily levels between 2013 and 2014 from both the original and binned data. Results are shown in Figures 3 and 4. In the figure legend, '1' (in Blue) indicates the links where activities increased when using the *original* dataset but decreased when using the *binned* dataset; '-1'(in Orange) indicates the links where activities decreased in the original data but where opposite results were obtained from the *binned* dataset. Lastly, '0' (in grey) shows the links where the results from the *binned* data set match with the results from the *original* dataset.

From the figures we can observe that *binned* dataset not only loses information but can also provide directly contradictory results. For the hourly aggregation, a considerable number of links show a decrease in activity with *binned* data while it shows an increase with the *original* data. The binning criteria will round-down more activities if the activities are spread across the day uniformly whereas it will round-up the activities if the activities are bunched together for particular hours. Even though activities increased for the year 2014, the activities are evenly distributed across the day and thus are lost due to the binning criteria whereas for the year 2013, these activities were bunched together and hence were not lost after binning. The difference in change in activities between the *original* dataset and the *binned* dataset can be as high as a few hundred activities on these links. For the daily aggregation, the number of links in the '1' category reduces but increases in the number of links in the '-1' category are more than compared to results from the hourly aggregation. This is due to the rounding-up error in the daily aggregation and thus links show that the activities have increased for the *binned* dataset whereas in *original* dataset these values have reduced. This finding is crucial for studies following a similar methodology to Hochmair et al., (2019); Sun & Mobasheri, (2017), that are investigating differences in commuting and non-commuting trips; as commuting trips are more likely to be bunched together (peak periods) whereas recreational trips can be uniformly spread across a given day.

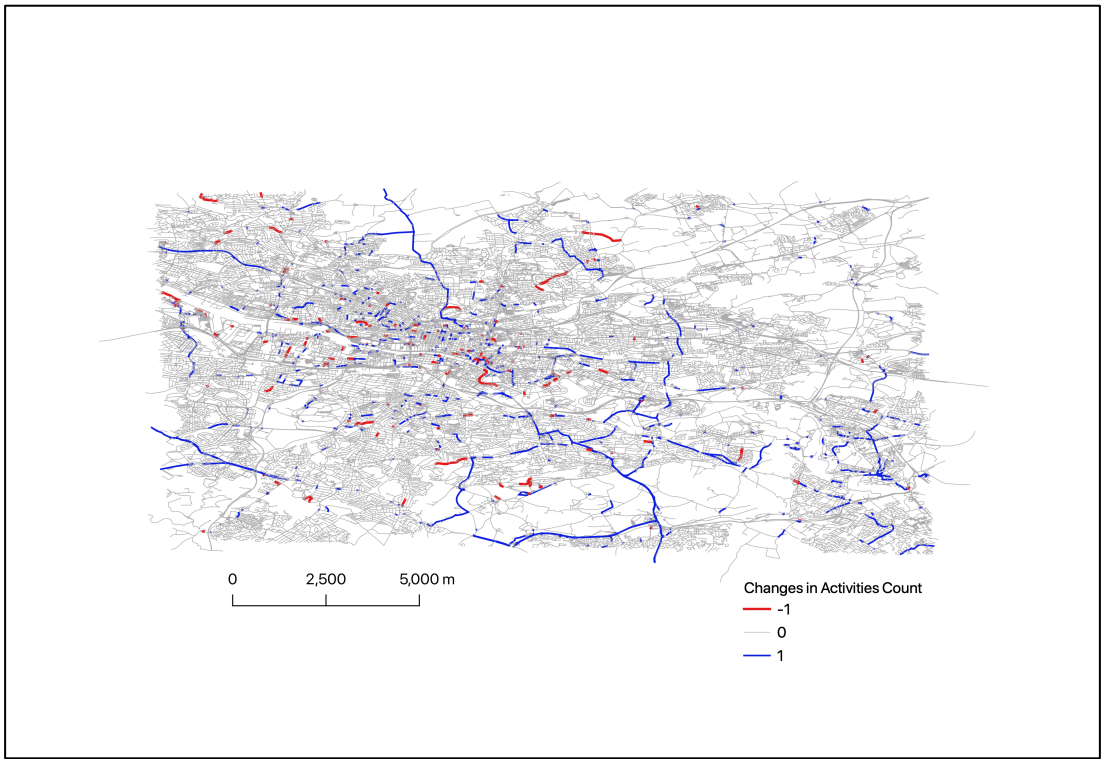

Figure 3: Change in Activities count (Hourly aggregation)

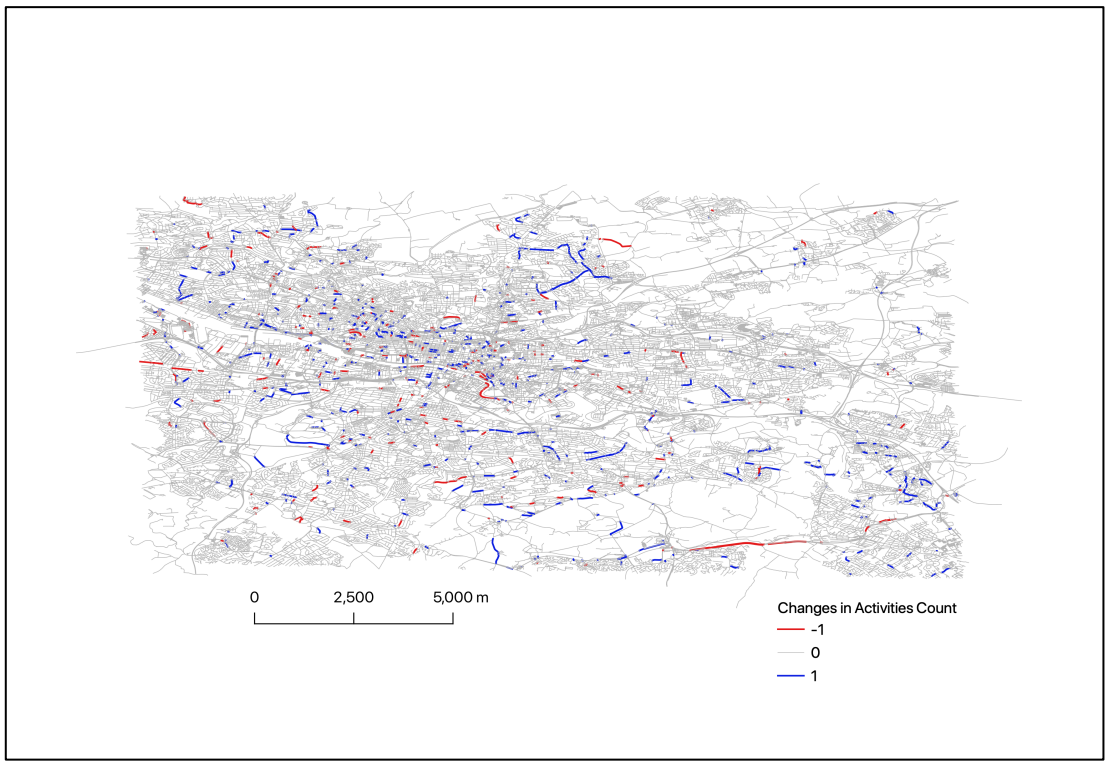

Figure 4: Change in Activities count (Daily aggregation)

4.4 RQ4) To what extent do the research outcomes of cycling infrastructure impact analysis change due to the binning strategy?

We formulated the weight matrix for the Glasgow city network using the Geoda software, and the connectivity histogram is shown in the Figure 5.

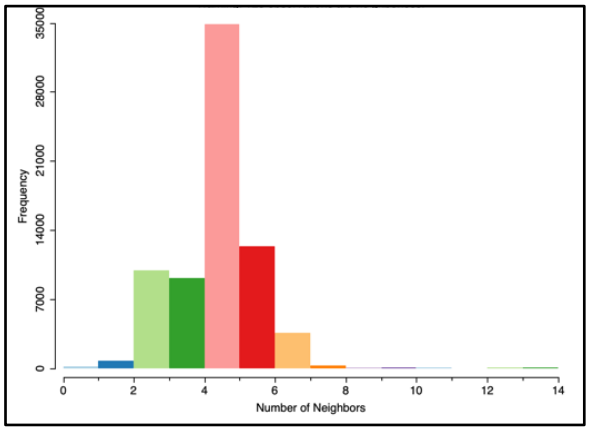

Figure 5: Connectivity Histogram

A quick glance at the histogram informs us that majority of the links have 4 neighbours (mostly intersections), and significant number of links have neighbours in the range of 2 to 6. There are few links with high numbers of neighbours and there are 128 links with no neighbours (isolates). These neighbourless links or isolates are links which are mostly at the outer edges of the shape file and are ignored by the software for further analysis.

We applied LISA to the change in activities for both the *original* and the *binned* datasets. Interestingly, the spatial autocorrelation results for the original dataset for all aggregations are identical. Thus, we provide only one result as a common one for the *original* dataset shown in Figure 6. Figures 7 and 8 show the spatial autocorrelation results for the change in activities for the *binned* dataset with hourly aggregation and daily aggregation respectively. Here, the High-High clusters (links in red) are where there is a significant increase activity on links surrounded by other links with high increase in activity. Low-Low clusters (links in dark blue) are where there is a significant decrease in activity on a link surrounded by other links with reduced activity. Similarly, Low-High (links in light blue) are the links where there is reduction in activities but with neighbours with increased activities. These are generally smaller roads that are connected to the major routes. Finally, High-Low (links in pink) are the clusters where links with increased activity are surrounded by links with decreased activity.

Even at a high significance level ($p = 0.01$), we observe that results for the *binned* data for hourly aggregation vary significantly from the results of the *original* dataset. In the *binned* data with hourly aggregation, looking into Low-Low clusters, we notice that certain links show a statistically significant reduction in activities whereas results from the *original* dataset show that these links are not statistically significant. As determined by the results of RQ3, these are the same links where activities for the original dataset increased rather than decrease. Therefore, these results indicate that we not only lose information from the dataset due to these binning criteria but can also produce opposite conclusions.

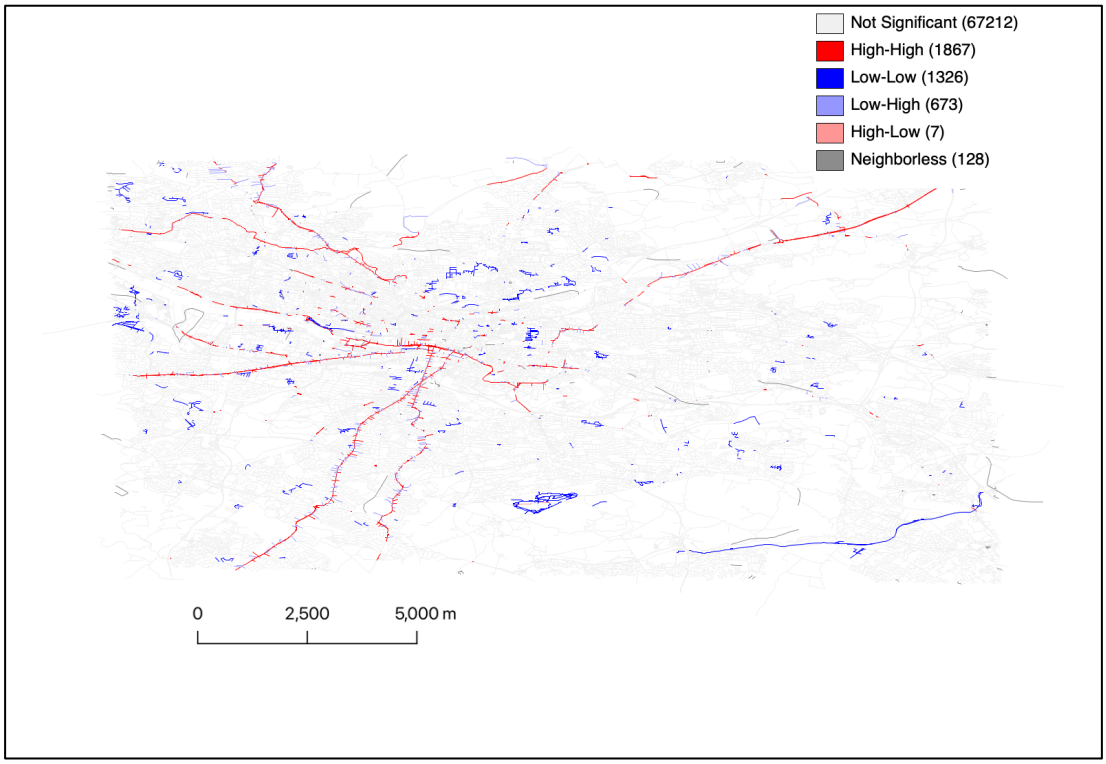

Figure 6: Spatial Autocorrelation results for change in activities (Original Dataset: Common)

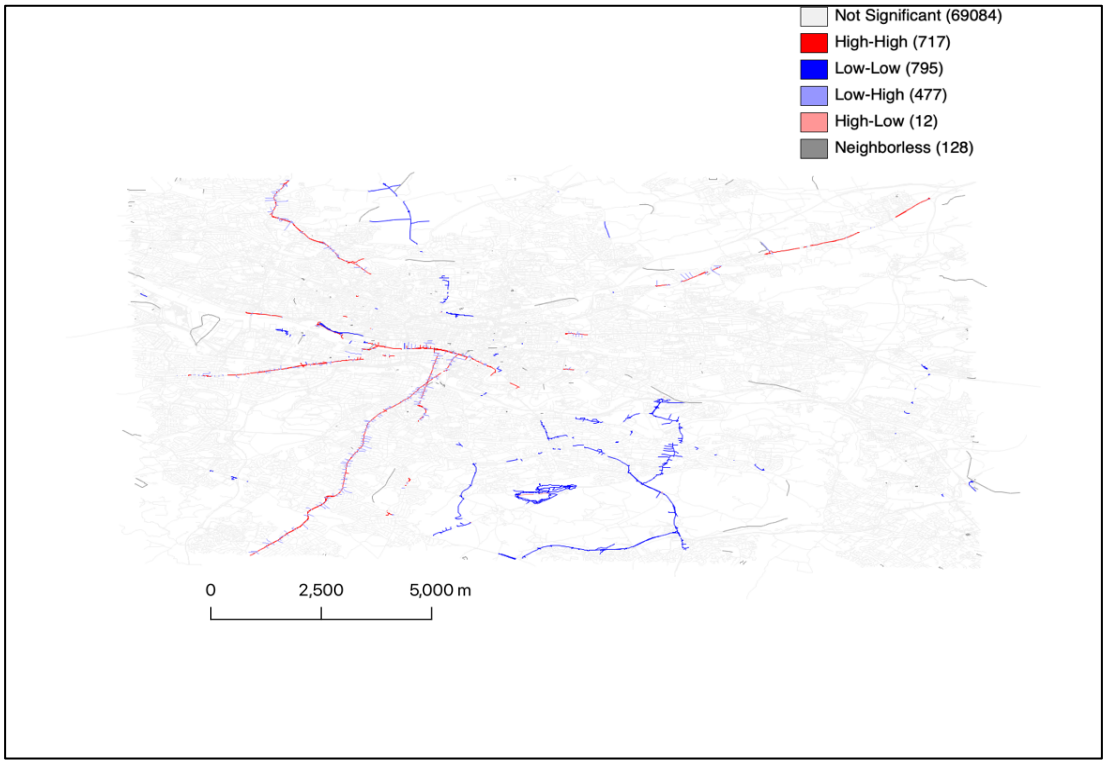

Figure 7: Spatial Autocorrelation results for change in activities (Binned Dataset: Hourly Aggregation)

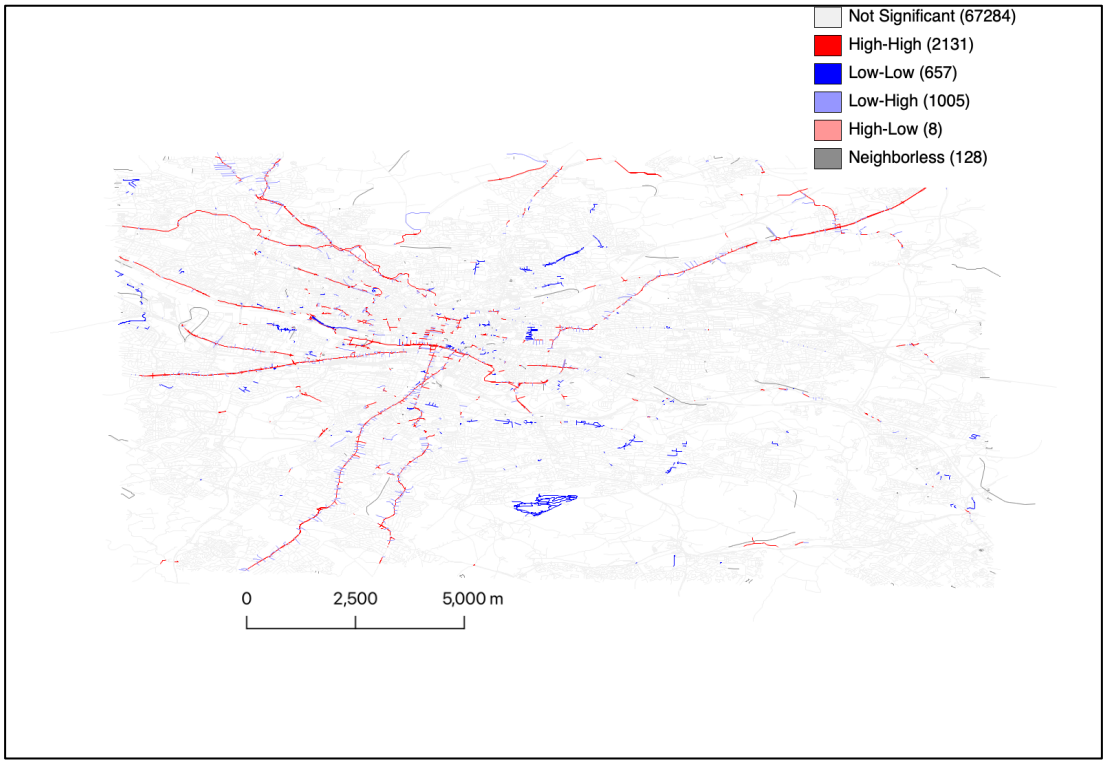

Figure 8: Spatial Autocorrelation results for change in activities (Binned Dataset: Daily Aggregation)

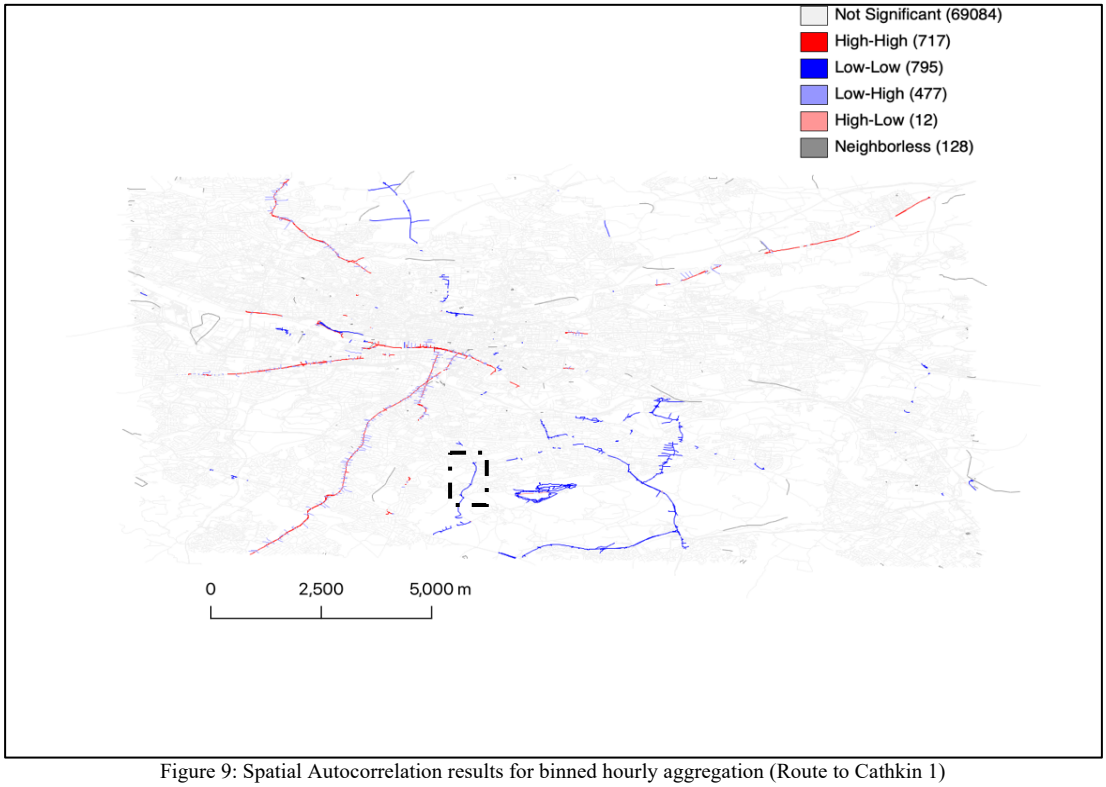

Figure 9: Spatial Autocorrelation results for binned hourly aggregation (Route to Cathkin 1)

A closer look on the spatial autocorrelation results from the *binned* hourly aggregated dataset highlights that a part of the Cathkin 1 infrastructure (marked in dashed box in Figure 9) lies on the links showing a significant reduction in activity. The results from the *original* Strava dataset (Figure 6) show that the changes (i.e., increase in the total cycling activities made by Strava users) are random and no spatial dependence is observed whereas the results from the *binned* hourly aggregated dataset would have concluded that the infrastructure development has caused a reduction in cycling activities. Figure 10 shows the part of the infrastructure on the spatial autocorrelation results for binned hourly aggregation. These results raise serious concerns whether these datasets can be used at these finer temporal aggregations not only for infrastructure analysis but also for analysis like peak and off-peak travel behaviour.

5. Discussion and Conclusion

As a result of the limitations of the data collected by traditional sources for analysing active travel behaviour, crowdsourced data are becoming increasingly popular. New forms of data provide various opportunities and have transformational potential for research. Rising apprehensions over individual privacy and data protection has led to the adoption of various methods of anonymising. A common approach is to apply some form of binning criteria, where data are gathered in bins to protect users' identities. By doing so, data loses information and consequently the utility of the data is reduced.

In this paper, we investigated the number of activities lost due to the binning criteria applied by Strava Metro as an example. We compare ground truth data in the form of a cordon count survey to both the *original* Strava data and the *binned* data set. We demonstrate the loss of power of prediction when using the *binned* data compared the results with the *original* Strava data. We have also calculated the error due to the Strava binning criteria at these different temporal aggregations and found that the net error (derived by comparing under and overcounting due to rounding) can be as high as 90% for hourly aggregations and around 50% for daily aggregations. The net error reduces for monthly aggregations. We also found that these net errors reduced where the volume of cycling activity increases. Where there are fewer Strava users and bicycle activities these errors may be increased. These errors will increase further if applied on the directed counts instead of total counts on the link and might render the dataset impractical to use for certain policy analyses. As mentioned in the literature section, Conrow et al., (2018) used directional flows to compare manual count data with Strava data. With net errors expected to be higher for flows segregated by direction, even at larger temporal aggregation, usage of this data may become ineffective. Binning criteria aiming to minimise information loss can assist in improving the quality of the dataset while maintaining the privacy of the users.

By analysing the change in activities between consecutive years, we saw that the results from the *binned* dataset deviated significantly from the *original* dataset's results. Some links even showed opposite results while capturing these changes due to differences in the temporal distribution of these activities. These results indicate that *binned* dataset could give erroneous results especially in the case of examining commuter and non-commuters. Our results show that the utility of *binned* Strava data reduces due to the activity count loss compared to the *original* Strava data. This implies that the outcomes of cycling research could change when using *binned* Strava data, potentially resulting in spurious conclusions.

Lastly, conducting spatial analysis on the change in activities suggested that using finer temporal aggregation can lead to contrasting conclusions especially in the context of infrastructure assessment and thus impact research outcomes. This raises serious concerns about the quality of the *binned* dataset for research purposes especially in the transport geography and policy fields. Using monthly aggregated data can resolve this issue but then constraints the overall usability of the dataset. These results highlight the importance of understanding the impacts of binning on the quality of the data.

This research raises serious concerns about the use of the Strava Metro product for certain types of analysis. We have demonstrated how the binning strategy, introduced by Strava to safeguard their users' privacy, significantly impacts on the product's functionality. Our research has important implications for users of Strava Metro data. The new binned Strava Metro product was introduced in April of 2018. This may have implications for the results of any published paper which used the new product, especially if they contain analysis where numbers of cyclists may be low and where they use data that has not been aggregated for over a month. Researchers who are currently working on these data will need to account for how they deal with the current issues caused by Strava's binning strategy. Results from papers already published, that indicate they use *binned* Strava data should be treated with caution. There are a number of papers that have been published exploring the accuracy and validity of Strava data in a number of settings which use the un-binned data (Griffin & Jiao, 2015; Jestico et al., 2016). Many of these papers highlight how the data can be used in research, with certain caveats, however, these papers are not relevant to the use of the new Strava Metro product as the binned data does not have the same level of precision as the previous Strava product.

The original Strava Metro product has been shown to be a valuable tool for academics who need reliable cycling data if they are to understand factors like the impact of new policies, infrastructure, exposure in cycling casualty studies (Hong et al., 2019; Boss et al., 2018). The new product, while it may offer some functionality for local authorities to identify popular cycling routes, has less to offer academics where accuracy is vital to adding to our understanding. We recommend that Strava consider a strategy which protects the privacy of individuals while preserving the integrity of the data.

Extending this work to other forms of data is a possible direction for future work. Various rounding strategies such as random rounding, controlled rounding or semi-controlled rounding etc. have been discussed in the handbook of statistical disclosure (Hundepool et. al, 2010) and can be viewed as an alternative to the conventional rounding of Strava. Further research can determine the binning criteria which minimises information loss, hence maintaining the quality of the dataset while still ensuring that the privacy of individuals is protected. Overall, this study highlights the impacts of the additional limitations placed on the dataset by the providers. These additional limitations must be taken into account by researchers intending to utilise this data in future.

Author Contributions:

Varun Raturi: Conceptualization, Methodology, Writing- Original draft preparation, Data Curation, Software; **Jinhyun Hong:** Conceptualization, Methodology, Supervision, Writing- Reviewing and Editing; **David Philip McArthur:** Conceptualization, Supervision, Writing- Reviewing and Editing; **Mark Livingston:** Writing- Reviewing and Editing

Acknowledgments: The authors would like to acknowledge support from the Economic and Social Research Council-funded Urban Big Data Centre at the University of Glasgow (Grant ES/L011921/1, ES/S007105/1).

References

Annual Cycling Monitoring Report, Cycling Scotland. (2018), Retrieved from https://www.cycling.scot/mediaLibrary/other/english/3028.pdf

Anselin, L. (1995). Local Indicators of Spatial Association—LISA. *Geographical Analysis*, *27*(2), 93–115. https://doi.org/10.1111/j.1538-4632.1995.tb00338.x

Boss, D., Nelson, T., Winters, M., & Ferster, C. J. (2018). Using crowdsourced data to monitor change in spatial patterns of bicycle ridership. *Journal of Transport and Health*, *9*(February), 226–233. https://doi.org/10.1016/j.jth.2018.02.008

Campbell, M. J., Dennison, P. E., Butler, B. W., & Page, W. G. (2019). Using crowdsourced fitness tracker data to model the relationship between slope and travel rates. *Applied Geography*, *106*(April), 93–107. https://doi.org/10.1016/j.apgeog.2019.03.008

Cavill, N., Kahlmeier, S., Rutter, H., Racioppi, F., & Oja, P. (2008). Economic analyses of transport infrastructure and policies including health effects related to cycling and walking: A systematic review. *Transport Policy*, *15*(5), 291–304. https://doi.org/10.1016/j.tranpol.2008.11.001

Celis-Morales, C. A., Lyall, D. M., Welsh, P., Anderson, J., Steell, L., Guo, Y., … Gill, J. M. R. (2017). Association between active commuting and incident cardiovascular disease, cancer, and mortality: prospective cohort study. *BMJ (Clinical Research Ed.)*, *357*, j1456. https://doi.org/10.1136/bmj.j1456

Chapman, R., Keall, M., Howden-Chapman, P., Grams, M., Witten, K., Randal, E., & Woodward, A. (2018). A cost benefit analysis of an active travel intervention with health and carbon emission reduction benefits. *International Journal of Environmental Research and Public Health*, *15*(5), 1–10. https://doi.org/10.3390/ijerph15050962

Chen, C., Anderson, J. C., Wang, H., Wang, Y., Vogt, R., & Hernandez, S. (2017). How bicycle level of traffic stress correlate with reported cyclist accidents injury severities: A geospatial and mixed logit analysis. *Accident Analysis and Prevention*, *108*(May), 234–244. https://doi.org/10.1016/j.aap.2017.09.001

Claire Welsh, Carlos A Celis-Morales, Frederick Ho, Donald M Lyall, Daniel Mackay, Lyn Ferguson, Naveed Sattar, Stuart R Gray, Jason M R Gill, Jill P Pell, Paul Welsh. *Association of injury related hospital admissions with commuting by bicycle in the UK: prospective population based study. BMJ, 2020*; m336 DOI: 10.1136/bmj.m336

Conrow, L., Wentz, E., Nelson, T., & Pettit, C. (2018). Comparing spatial patterns of crowdsourced and conventional bicycling datasets. *Applied Geography*, *92*(January), 21–30. https://doi.org/10.1016/j.apgeog.2018.01.009

de Hartog, J. J., Boogaard, H., Nijland, H., & Hoek, G. (2010). Do the health benefits of cycling outweigh the risks? *Environmental Health Perspectives*, *118*(8), 1109–1116. https://doi.org/10.1289/ehp.0901747

Eisenman, S. B., Miluzzo, E., Lane, N. D., Peterson, R. A., Ahn, G. S., & Campbell, A. T. (2009). BikeNet: A mobile sensing system for cyclist experience mapping. *ACM Transactions on Sensor Networks*, *6*(1). https://doi.org/10.1145/1653760.1653766

El Esawey, M., Mosa, A. I., & Nasr, K. (2015). Estimation of daily bicycle traffic volumes using sparse data. *Computers, Environment and Urban Systems*, *54*, 195–203. https://doi.org/10.1016/j.compenvurbsys.2015.09.002

Froehlich, J., Neumann, J., & Oliver, N. (2009). Sensing and predicting the pulse of the city through shared bicycling. *IJCAI International Joint Conference on Artificial Intelligence*, (3), 1420–1426.

Glasgow City Council. (2016). *Glasgow's Strategic Plan for Cycling 2016 - 2025*. 60. Retrieved from https://www.google.com/search?safe=active&biw=1600&bih=757&sxsrf=ACYBGNQOy_PFrlpoCsJX3Pa6PK1HdWURUw%3A1575547952650&ei=MPToXcKZJ4aZlwTR7LXgBg&q=Glasgow%27s+Strategic+Plan+for+Cycling+2016+-+2025&oq=Glasgow%27s+Strategic+Plan+for+Cycling+2016+-+2025&gs_l=p

Griffin, G. P., & Jiao, J. (2015). Where does bicycling for health happen? Analysing volunteered geographic information through place and plexus. *Journal of Transport and Health*, *2*(2), 238–247. https://doi.org/10.1016/j.jth.2014.12.001

Harvey, F., & Krizek, K. (2007). Commuter Bicyclist Behavior and Facility Disruption. *Transportation Research Board*, 60. Retrieved from https://trid.trb.org/view.aspx?id=811576

Haworth, J. (2016). Investigating the potential of activity tracking app data to estimate cycle flows in urban areas. The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences - XXIII ISPRS Congress. Prague, Czech Republic: Commission II, WG II/8.

Hochmair, H. H., Bardin, E., & Ahmouda, A. (2019). Estimating bicycle trip volume for Miami-Dade county from Strava tracking data. *Journal of Transport Geography*, *75*(January), 58–69. https://doi.org/10.1016/j.jtrangeo.2019.01.013

Hong, J., McArthur, D. P., & Livingston, M. (2020). The evaluation of large cycling infrastructure investments in Glasgow using crowdsourced cycle data. *Transportation*, (0123456789). https://doi.org/10.1007/s11116-019-09988-4

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Lenz, R., Nylor, J., Schulte Nordholt, E., Seri, G., & De Wolf, P.-P. (2010). Handbook on statistical disclosure control. http://neon.vb.cbs.nl/casc/.%5CSDC_Handbook.pdf. Retrieved 10 Nov 2010.

Jestico, B., Nelson, T., & Winters, M. (2016). Mapping ridership using crowdsourced cycling data. *Journal of Transport Geography*, *52*, 90–97. https://doi.org/10.1016/j.jtrangeo.2016.03.006

Keler, A., & Mazimpaka, J. D. (2016). Safety-aware routing for motorised tourists based on open data and VGI. *Journal of Location Based Services*, *10*(1), 64–77. https://doi.org/10.1080/17489725.2016.1170216

Kitchin, R. (Maynooth U. (2016). *Getting Smarter About Smart Cities: Improving Data Privacy and Data Security*. 82. Retrieved from http://eprints.maynoothuniversity.ie/7242/1/Smart

Lederman, J., Taylor, B. D., & Garrett, M. (2016). A private matter: the implications of privacy regulations for intelligent transportation systems. *Transportation Planning and Technology*, *39*(2), 115–135. https://doi.org/10.1080/03081060.2015.1127537

Lee, K., & Sener, I. N. (2019). Understanding potential exposure of bicyclists on roadways to traffic-related air pollution: Findings from El Paso, Texas, using Strava metro data. *International Journal of Environmental Research and Public Health*, *16*(3). https://doi.org/10.3390/ijerph16030371

Lin, Z., Hewett, M., & Altman, R. B. (2002). Using binning to maintain confidentiality of medical data. *Proceedings / AMIA ... Annual Symposium. AMIA Symposium*, 454–458.

Livingston, M., McArthur, D., Hong, J., & English, K. (2020). Predicting cycling volumes using crowdsourced activity data. *Environment and Planning B: Urban Analytics and City Science*, 239980832092582. https://doi.org/10.1177/2399808320925822

McArthur, D. P., & Hong, J. (2019). Visualising where commuting cyclists travel using crowdsourced data. *Journal of Transport Geography*, *74*(December 2018), 233–241. https://doi.org/10.1016/j.jtrangeo.2018.11.018

Mindell, J. S. (2015). Active travel is (generally) good for health, the environment and the economy. *Journal of Transport and Health*, *2*(4), 447–448. https://doi.org/10.1016/j.jth.2015.10.006

Natalie Mueller, David Rojas-Rueda, Tom Cole-Hunter, Audrey de Nazelle, Evi Dons, Regine Gerike, Mark Nieuwenhuijsen (2015). Health impact assessment of active transportation: A systematic review. *Preventive Medicine*. Retrieved from https://pdf.sciencedirectassets.com/272375/1-s2.0-S0091743515X00065/1-s2.0-S0091743515001164/main.pdf

Norman, P., Pickering, C. M., & Castley, G. (2019). What can volunteered geographic information tell us about the different ways mountain bikers, runners and walkers use urban reserves? *Landscape and Urban Planning*, *185*(February), 180–190.

https://doi.org/10.1016/j.landurbplan.2019.02.015
Nunan, D., & Di Domenico, M. (2017). Big Data: A Normal Accident Waiting to Happen? *Journal of Business Ethics*, *145*(3), 481–491. https://doi.org/10.1007/s10551-015-2904-x
Oja, P., Titze, S., Bauman, A., de Geus, B., Krenn, P., Reger-Nash, B., & Kohlberger, T. (2011). Health benefits of cycling: A systematic review. *Scandinavian Journal of Medicine and Science in Sports*, *21*(4), 496–509. https://doi.org/10.1111/j.1600-0838.2011.01299.x
Orellana, D., & Guerrero, M. L. (2019). Exploring the influence of road network structure on the spatial behaviour of cyclists using crowdsourced data. *Environment and Planning B: Urban Analytics and City Science*, *46*(7), 1314–1330. https://doi.org/10.1177/2399808319863810
Padgham, M. (2012). Human movement is both diffusive and directed. *PLoS ONE*, *7*(5), 1–11. https://doi.org/10.1371/journal.pone.0037754
Perkins, T. & Blake, D. (2016). Understanding Cycling Behaviour in Boomtown Perth. In: Biermann, s., Olaru, d. & Paül, v. (eds.) Planning Boomtown and Beyond. UWA Publishing.
Prandi, F., Soave, M., Devigili, F., & Amicis, R. De. (2014). Collaboratively collected geodata to support routing service for disabled people. *Proceedings of the 11th International Symposium on Location-Based Services*, (November), 26–28. https://doi.org/10.13140/2.1.2937.1203
Proulx, F. R., & Pozdnukhov, A. (2017). Bicycle Traffic Volume Estimation using Geographically Weighted Data Fusion. *Journal of Transport Geography*, 1–14.
Rissel, C. E. (2009). Active travel: a climate change mitigation strategy with co-benefits for health. *New South Wales Public Health Bulletin*, *20*(1–2), 10–13. https://doi.org/10.1071/nb08043
Romanillos, G., Zaltz Austwick, M., Ettema, D., & De Kruijf, J. (2016). Big Data and Cycling. *Transport Reviews*, *36*(1), 114–133. https://doi.org/10.1080/01441647.2015.1084067
Roy, A., Nelson, T. A., Fotheringham, A. S., & Winters, M. (2019). Correcting Bias in Crowdsourced Data to Map Bicycle Ridership of All Bicyclists. *Urban Science*, *3*(2), 62. https://doi.org/10.3390/urbansci3020062
Roy, J. R., Batten, D. F., & Lesse, P. F. (1982). Minimizing Information Loss in Simple Aggregation. *Environment and Planning A: Economy and Space*, *14*(7), 973–980. https://doi.org/10.1068/a140973
Rubinstein, I. S. (2013). Big data: The end of privacy or a new beginning? *International Data Privacy Law*, *3*(2), 74–87. https://doi.org/10.1093/idpl/ips036
Saad, M., Abdel-Aty, M., Lee, J., & Cai, Q. (2019). Bicycle Safety Analysis at Intersections from Crowdsourced Data. *Transportation Research Record*, *2673*(4), 1–14. https://doi.org/10.1177/0361198119836764
Saha, D., Alluri, P., Gan, A., & Wu, W. (2018). Spatial analysis of macro-level bicycle crashes using the class of conditional autoregressive models. *Accident Analysis and Prevention*, *118*(July 2017), 166–177. https://doi.org/10.1016/j.aap.2018.02.014
Sanders, R. L., Frackelton, A., Gardner, S., Schneider, R., & Hintze, M. (2017). Ballpark method for estimating pedestrian and bicyclist exposure in Seattle, Washington: Potential option for resource-constrained cities in an age of big data. *Transportation Research Record*, *2605*(1), 32–44. https://doi.org/10.3141/2605-03
Stopher, P., FitzGerald, C., & Zhang, J. (2008). Deducing Mode and Purpose from GPS Data. *87th Annual Meeting of the Transportation Research Board*, 1–11.
Sun, Y., & Mobasheri, A. (2017). Utilizing crowdsourced data for studies of cycling and air pollution exposure: A case study using strava data. *International Journal of Environmental Research and Public Health*, *14*(3). https://doi.org/10.3390/ijerph14030274
Tene, O., & Polonetsky, J. (2012). Privacy in the Age of Big Data: A Time for Big Decisions. *Stanford Law Review Online*, *64*, 63. https://doi.org/10.5121/ijgca.2012.3203
Tobler, W. R. (1965). Computation of the correspondence of geographical patterns. *Papers of the Regional Science Association*, *15*(1), 131-139. https://doi.org/10.1007/BF01947869.
Wang, Y., Monsere, C. M., Chen, C., & Wang, H. (2018). Development of a Crash Risk-Scoring Tool for Pedestrian and Bicycle Projects in Oregon. *Transportation Research Record*, *2672*(32), 30–39. https://doi.org/10.1177/0361198118794285
Woodcock, J., Edwards, P., Tonne, C., Armstrong, B. G., Ashiru, O., Banister, D., … Roberts, I. (2009). Public health benefits of strategies to reduce greenhouse-gas emissions: urban land transport. *The Lancet*, *374*(9705), 1930–1943. https://doi.org/10.1016/S0140-6736(09)61714-1