Understanding the socio-demographic representation of Tamoco mobile phone app data in Glasgow City Region

Michael Sinclair[1], Saeed Maadi[2], Qunshan Zhao[3], Jinyun Hong[4], Nick Bailey[5]
[1,2,3,4,5]Urban Big Data Centre, School of Social and Political Science, University of Glasgow, UK

March 8, 2022

Summary

This research advances the home location detection of new forms of mobile phone data by using high-resolution land-use data and compares resulting home location estimates with public socio-demographic data. The research allows us to identify potential biases in mobile phone data which may arise through uneven population coverage. Results show that the number of mobile phone users estimated across the Glasgow City Region in 2020 are proportional to the working population across different socio-demographic groups based on the Scottish Index of Multiple Deprivation.

KEYWORDS: spatial big data; mobile phone app data; home location detection; socio-demographic representativeness

1. Introduction

New forms of mobile phone (MP) data offer enormous potential by virtue of the volume of data available and the spatio-temporal details provided. However, the processes by which these data are produced are often rather unclear and they may also contain biases in population coverage which impact on the results they provide. An important dimension of applying spatial big data for analysing patterns of movements in the urban context concerns the characterization of where residents come from. Where someone resides is fundamental to understand how they use urban space and can be used to compare the representability of the data to that of the population.

The objective of this research is to explore the socio-demographic representativeness of new forms of MP data. We do this by advancing home location detection of new forms of MP, using high-resolution land use data, and comparing resulting estimates on home location to public socio-demographic data from the Scottish Index of Multiple Deprivation for Glasgow City and the seven surrounding administrative areas for 2020.

2. Data and methods

2.1. Data

The core data source for this research is a MP application dataset from Tamoco (https://www.tamoco.com/) which is available to the Urban Big Data Centre (UBDC, http://www.ubdc.ac.uk/) for non-commercial academic research. The data is an example of Location-Based Service data which is generated when a MP application updates the location of a mobile device using the most accurate available location sensor including single or a combination of GPS, Bluetooth, cellular tower signals, or Wi-Fi (Wang and Chen, 2018). This type of MP application data provides the point locations of the device with a certain degree of error (typically 10s of metres). However, in most cases, these data have a much higher spatial precision than other types of MP data such as call detail records which have been available for a longer period of time (Wang et al., 2018). The data is derived from user location data collected from a range of partner applications under an informed consent basis, limited to those age 16+. The dataset has personal identifiers replaced with non-reversible hashed identifiers which are changed monthly. The data has spatial coverage across the Glasgow City Region (eight contiguous local authorities), presented in Figure 1, with a population over 1.8 million people and the data is available for the calendar year 2020.

Datazones boundaries represent the key geography for the dissemination of small area statistics in

[1]Michael.sinclair@glasgow.ac.uk
[2]Saeed.maadi@glasgow.ac.uk
[3]Qunshan.zhao@glasgow.ac.uk
[4]Jinyun.hong@glasgow.ac.uk
[5]Nick.bailey@glasgow.ac.uk

Scotland (equivalent to Lower Super Output Areas in England) and are openly available. They are used for the analysis of home locations and to assign socio-demographic data. They are composed of groups of Census Output Areas and with a population in 2020 generally between 500-1000. In there study area there are over 2300 datazones.

Geomni 'UKBuildings' (created and maintained by Geomni, a Verisk company) land use data is utilised for the enrichment of the mobile data during the process of home location detection. The dataset represents the structure, characteristics, and use of commercial, public and residential buildings across the UK.

Scottish Index of Multiple Deprivation 2020 (SIMD) is utilized to identify socio-economic status (in terms of quintile, decile and percentile) at Datazone level. These are open data from the Scottish Government.

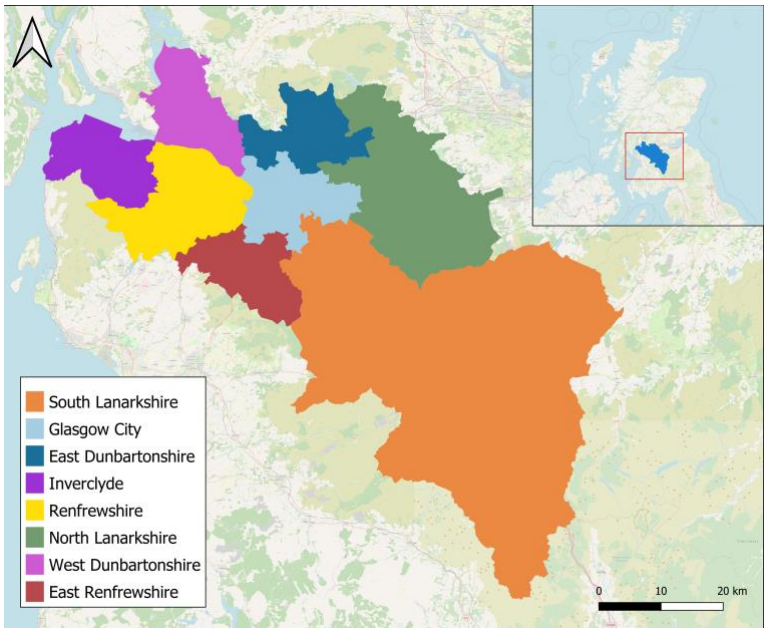

Figure 1 Study area.

2.2. Methods

The first step in the process of assigning aggregated socio-demographic data from SIMD is to identify the most likely home Datazone of each MP, i.e the area where someone is estimated to live. The most common home detection algorithms using MP data and other GPS data apply activity heuristics approaches (Alexander et al., 2015; Bojic et al., 2015; Pappalardo et al., 2021; Sinclair et al., 2020; Vanhoof et al., 2018; Wang et al., 2019), with the basic assumption that we can predict home location for an individual device based on the locations where activity concentrates during the night time. One major issue with this assumption is that evening work and leisure activities can lead to false estimates. Our methodology advances the state of the art by first enriching the MP data with high-resolution land use data from Geomni's UKBuildings layer before applying an activity heuristic approach based no nighttime activity. Among other improvements, this removes night-time activity related to non-residential locations and therefore more likely associated with night-time adult or leisure, and discounts these as home locations.

The home Datazone for an individual MP is estimated monthly and assumed as the Datazone which maximises the number of active evenings in residential and mixed residential space, where an evening is considered to be between 8 pm and 6 am. Only home locations returned with at least two active evenings for the same user in such space were considered for further analysis. In cases where more than one potential Datazone was returned in a given month, the user was not assigned a home location. Once a home Datazone was allocated, the user was assigned aggregate socio-demographic data from SIMD based on that Datazone. Data assigned included SIMD quintile, decile and percentile.

3. Results

In 2020 there were 1,023,098 unique user IDs in the Tamoco mobile phone dataset active in Glasgow City region (Figure 1). Since the user ID is changed monthly, a home location was estimated for all unique IDs though we can expect the same user to exist in the data multiple times and actual number of users to be lower the total unique IDs. Users generated 808,056,047 data points within the study area during 2020. Using the home detection method outlined in section 2.2 we were able to estimate a home Datazone for 244,263 user IDs (23.9%). The other users either: had only one active evening in residential or mixed residential space (7.6%); did not have data points falling within residential space or returned more than one potential home region (68.5%). Analysing the volume of data generated by users who could be assigned a home location we found that these 23.9% of user IDs were responsible for generating 81.7% of the data in 2020 (659,857,805 data points). This substantially strengthens our results and highlights that a large portion of users are responsible for generating a small portion of the data. This result also shows that without implementing the additional step of assigning data to residential space before estimating home location, we would likely incorrectly estimate home location for many of the other users.

Figure 2 presents the comparison between the percentage of mobile phone users and the percentage of the adult population across the eight councils of the Glasgow City Region (figure 1). Glasgow City and Inverclyde are slightly under-represented, while West Dunbartonshire and East Renfrewshire are slightly over-represented. Figure 3 presents the comparison between mobile and adult population within different SIMD quintile and decile groups. Results show that for both quintile and decile groups, the ratio of mobile phone users is very comparable to that of the adult population. Taking the decile level, although estimates vary slightly between months, the ratio between the average number of mobile users and the adult population ranges between 0.96 and 1.05 for different groups. That means the mobile population is only over or underrepresented by a maximum of 5% at the decile level when data is aggregated across the study area. Furthermore, at the percentile level, the results are also positive (Table 1). The correlation between mobile users and the adult population is 0.97 (p-value < 0.01) and this positive result holds across all regions, despite Inverclyde showing a slightly weaker correlation.

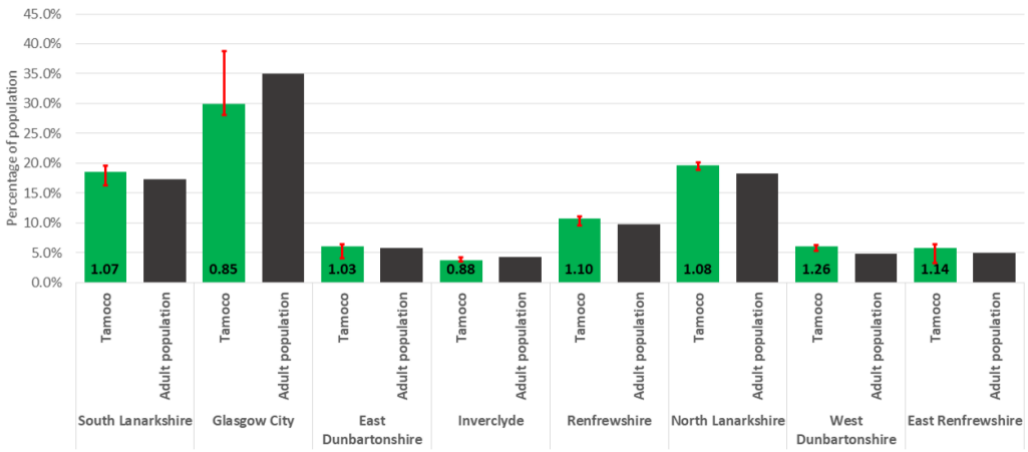

Figure 2 Geographic comparison between mobile phone users and adult population in 2020 for councils zones in the Glasgow City Region.

Notes: For council boundaries see figure 1; red error lines in A and B represent the range of results from individual months while the bar represents the mean number of mobile users per month. Labels are the ratio between the mean monthly mobile users and adult population in each region.

Table 1 Correlation results at SIMD percentile level between Tamoco mobile users and the adult population in 2020 for different councils in the Glasgow City Region

Year	Council	Percentiles included	Correlation
2020	South Lanarkshire	97	0.94
	Glasgow City	100	0.98
	East Dunbartonshire	59	0.94
	Inverclyde	63	0.87
	Renfrewshire	90	0.94
	North Lanarkshire	93	0.95
	West Dunbartonshire	64	0.94
	East Renfrewshire	58	0.97
	All regions	100	0.97

Notes: See figure 1 for council boundaries; results show Pearson's correlation coefficient; all correlations are significant (p-value <0.01); for Tamoco, the mean monthly users in a given year is used in the analysis (see methods for further details on the structure of the data); not all SIMD percentiles are present in each region.

Figure 3 Socio-demographic comparison between mobile phone users and adult population in 2020 at the quintile (A), decile (B) levels of the SIMD for the Glasgow City Region.

Notes: Red error lines in A and B represent the range of results from individual months while the bar represents the mean number of mobile users per month. Labels are the ratio between the mean monthly mobile users and adult population in each group.

4. Discussion and future work

In this research, we use MP application data from Tamoco to estimate the geographic coverage of MP users across one city-region in 2020 and compare results to the coverage of the adult population at different levels of the SIMD. The results are positive and show that the population from MP application data is comparable to that of the adult population across regions and different socio-demographic groups. The findings support the use of MP app data to further understand human activity in cities.

As a limitation, because the sample is restricted to data with the Glasgow City Region, it is unavoidable that some users will be incorrectly assigned to live within this area because we are missing part of their data. However, our approach using active residential evenings and limiting analysis to users with at least two active residential evenings in a month will likely reduce this error.

Future research can expand the analysis to multiple years and extend the comparison of socio-demographic results to include other products such as CACI's acorn consumer classification. Another area for future study with potential to make a large impact is to compare results between rival mobile phone datasets. While Glasgow City Region represents a substantial area and a large population, further research might also seek to expand the results to other cities in the UK.

Acknowledgements

The work was made possible by the ESRC's on-going support for the Urban Big Data Centre (UBDC) [ES/L011921/1 and ES/S007105/1].

References

Alexander L Jiang S Murga M and González M C (2015Origin–destination trips by purpose and time of day inferred from mobile phone data. Transportation Research Part C: Emerging Technologies. 58, 240–250. https://doi.org/10.1016/j.trc.2015.02.018

Bojic I Massaro E Belyi A Sobolevsky S and Ratti C (2015) Choosing the Right Home Location Definition Method for the Given Dataset, in: Liu, T.-Y., Scollon, C.N., Zhu, W. (Eds.), Social Informatics, Lecture Notes in Computer Science. Springer International Publishing, Cham, 194–208. https://doi.org/10.1007/978-3-319-27433-1_14

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

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

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

Wang F and 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 Z He S Y and Leung Y (2018) Applying mobile phone data to travel behaviour research: A literature review. Travel Behaviour and Society, 11, 141–155. https://doi.org/10.1016/j.tbs.2017.02.005

Wang F Wang J Cao J Chen C and Ban X (2019) Extracting trips from multi-sourced data for mobility

pattern analysis: An app-based data example. Transp. Res. Part C Emerg. Technol. 105, 183–202. https://doi.org/10.1016/j.trc.2019.05.028

Biographies

Michael Sinclair is an early career researcher with the Urban Big Data Centre. His research interests focus on using new forms of spatial data to understand human perceptions and behaviour for sustainability research.

Saeed Maadi is a research associate with the Urban Big Data Centre. His research focuses on urban transport and mobility using new forms of spatial big data.

Qunshan Zhao is a Lecturer in Urban Analytics in the Urban Big Data Centre at University of Glasgow. His research interests span urban analytics, spatial optimization/statistics, remote sensing, and geographic information science, with applications in transportation, housing, public health, and environmental monitoring.

Jinhyun Hong is a Senior Lecturer in Transportation Planning in the Urban Big Data Centre at University of Glasgow. His research interests include: Interaction among built environment; travel behaviour and air quality; ICT, transport and places; Transportation planning.

Nick Bailey is a Professor in Urban Studies and Director of the Urban Big Data Centre at University of Glasgow. Nick is interested in how neighbourhoods shape our lives and how divisions between richer and poorer areas within the city are created and maintained. He is also interested in broader issues of poverty and social exclusion.