Spatially aggregating mobility data – implications and challenges

Louise Sieg[*], James Cheshire[†]

Department of Geography, University College London

February 15, 2021

Summary

This paper explores the discrepancies in analysis resulting from aggregating data to different scales. We link mobile phone activity data to varying statistical geographies and compare the population impression obtained by joining these aggregates to a geodemographic classification (Workplace Zone Classification). We aim to provide an initial evaluation of how to approach creating aggregated products which preserve privacy but remain representative of the original dataset.

KEYWORDS: Geodemographics, Geospatial 'Big Data', Urban Analytics

Introduction

The emergence of new forms of data has changed the way we conduct research in human mobility. Mobile phone data is widely used to map and predict activity levels in urban centers and has been at the forefront of mobility research on COVID-19 (Zhou et. al, 2020) (Chang et. al, 2020). In parallel, the appearance of consumer data presents both prospects and challenges for research (Kitchin, 2013). One challenge is identifying what information is lost when the data is repurposed for research (Lansley and Cheshire, 2018).

De-identification of data is not sufficient to fully protect data privacy, and aggregates are a central aspect of data misuse prevention (de Montoje, 2018). However, granularity can be lost in the process. Furthermore, the creation of aggregates for third party use implies a wide range of future applications for the data sub-product, complicating decision-making.

Research aim

This paper seeks to explore the impacts of different levels of aggregation on the linkage of mobile phone activity data to standard statistical geographies and products. Whilst the modifiable areal unit problem (MAUP) will mean there is no perfect answer it is our hope that we can establish a pragmatic level of aggregation that preserves privacy whilst maintaining meaningful linkage. We first aggregate a granular location dataset using different scales and statistical boundaries before evaluating the discrepancies in population representations when these different aggregates are linked with geodemographic characteristics. This aims to outline the decision-making behind minimising the repercussions of aggregation and inform our future decisions regarding similar datasets.

Data

This paper focuses on the borough of Westminster. The timeframe is the working week of 16/09/19-20/09/19, over rush hours (8-10am).

1.1. Mobile phone location data

Huq Industries provided the mobile phone location data (https://huq.io/). The dataset consists of

[*] louise.sieg.16@ucl.ac.uk
[†] james.cheshire@ucl.ac.uk

devices' GPS location and timestamp stored by mobile applications (apps) when the app is being used (Trasberg, 2020). The apps seek user's consent to store this information. Though the original dataset consists of terabytes of data, the subset for this study is around 200,100 datapoints. It is stored securely in an ISO27001 facility with access restricted to researchers who have undertaken the Consumer Data Research Centre (CDRC) accreditation programme and following a multi-stage application process.

1.2. Workplace Zone Classification (WPZC)

The WPZC is a geodemographic classification used to describe the working populations and workplace geographies of London (Singleton et al, 2017). Each workplace zone within the classification is assigned a Group and Subgroup, the Subgroup being most relevant at the scale of our analysis. Population weighted centroids were used in the spatial linkage of WPZCs to mobile phone activity data.

1.3. Geographic areas and grids

Aggregating the specific locations generated by the phone apps to simple counts within a grid removes both the identifiers (phone IDs) and overly precise locations. Assuming a level of aggregation that ensures counts exceed a minimum level to ensure disclosure controls are met means that it is possible to derive an anonymised dataset that could be analysed more widely outside of a secure environment. However, it is not clear what impacts such aggregation would have on the effectiveness of linkage to other statistical data. To assess this, we create an aggregate by directly linking the data points to the WPZs geographies. Two categories of geographies are evaluated against it:

1. Ordnance Survey National Grids (OSGB). Three are used: 1km by 1km grid cells, 500x500m and 250x250m.

2. Output Areas (OA) at different scales. OA, Lower Layer Super Output-Areas (LSOA) and Medium Level Super Output-Areas (MSOA) are compared to OSGBs and the control. These typically describe residential ("night-time") statistics. While OAs are built around where different people *live*, WPZ focuses on where they work, creating imprecise results if WPZ were only linked to OAs.

Methodology

This methodology is built upon Trasberg and Cheshire's work, who previously used the Huq Industries dataset to create similar aggregates (Trasberg and Cheshire, 2020).

1. Measure the activity level in Westminster.

Activity levels are computed by counting the number of unique phone IDs per hour per grid cell. This process it repeated for each geographic area chosen, creating the aggregated products at each scale. For smaller scales, this implies obscuring activity counts below 10 to protect users. We selected the borough of Westminster for its varied WPZ subgroups and high activity counts from our dataset.

2. Link to WPZC.

Each grid cell (and OA scales) is assigned a WPZ Subgroup. To do this, we use population weighted centroids linked to workplace counts, as linking a Subgroup to a cell by area alone may not be representative of the underlying demographics. If more than one Subgroup overlaps a cell, the one with the highest weight (highest population count) is assigned to that cell.

3. Compare the geodemographic profiles obtained.

The assignment of WPZ to activity grids creates activity counts per Subgroup. This count varies depending on the scale at which the unique IDs were aggregated, and the activities created. We examine

how these vary from the control (direct join to WPZ geography) and inspect which one provide a different impression of the area's working demographics.

Results

Figure 1 visualises the different spatial distribution of WPZ Subgroups when joined at different scales (grey areas represent parks).

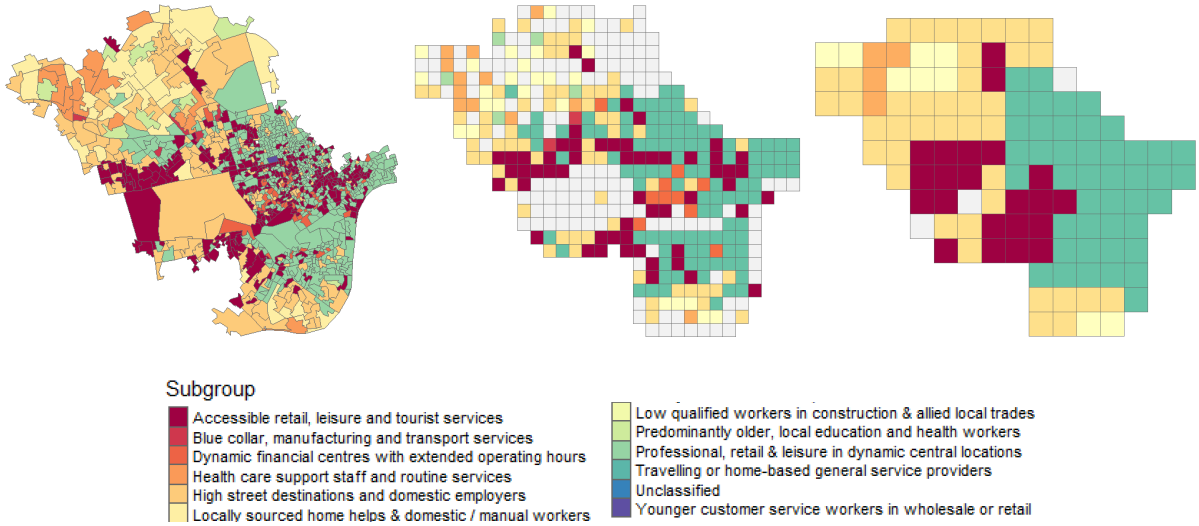

Figure 1 Westminster WPZ subgroups spatial distribution (WPZ, 250x250m and 500x500 joins).

Figure 2 shows the activity mean for the study period per aggregate per subgroup (left). As expected, aggregates present less overall activity than the control. Right, the percentage of aggregated activities allocated to each subgroup shows that 250x250m grid cell seems to be closest to the direct WPZ join for of population impression.

Figure 2: activity means (left) and Subgroup proportion (right).

Running a correlation test between the control activity counts and the aggregated ones revealed that, despite having removed low activity counts for OAs and 250x250m for anonymity purposes, smaller scales still present the highest correlation with control (respectively R=0.96 and 0.98). OAs do not seem to perform better than grids for day-time analysis concerning WPZCs.

Figure 3 shows the 250x250m Subgroup proportions against the control's, visualising which populations are over-estimated or under-estimated by the aggregated product. Where the color overfills the bar outline, the subgroup is over-estimated by the aggregate, when white remains it under-estimates. *Professional, retail & leisure* is over-estimated by ~10%. Discrepancies could be explained by population densities and dynamism varying between groups. Further analysis would be needed to account for these gaps.

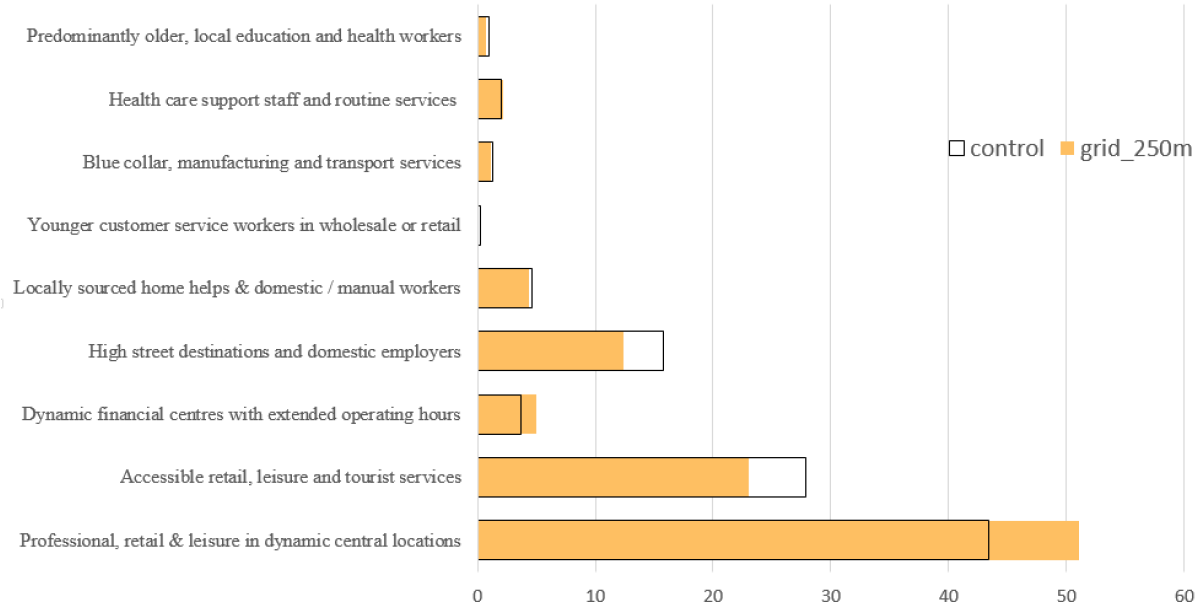

Figure 3 Ratio comparisons between control and 250x250m

Discussion

This paper aimed to highlight the discrepancies in analysis resulting from varying aggregation method. In doing so, it raises concerns of data bias when converting large consumer dataset into exploitable research data. Though aggregating points to 250x250 or OA, for instance, seemed to preserve part of the population impression, discrepancies in results cannot be neglected, especially as these may deepen when applied to larger scales. Density measures would have improved this analysis and provided potential solution to the rescaling of aggregates. However, this exploratory analysis suggests that grid cells are a pertinent starting point, should the data be used across analysis ranging from census to "daytime" statistics such as the WPZs. Moving forward, we hope to explore more versatile and accurate cell systems (such as ISEA3H) for more robust aggregates (Sahr et. al, 2003).

Acknowledgements

This research was funded by the ESRC. We thank Huq industries for providing data, and the UCL Research Ethics Committee for reviewing the project.

References

Chang, S., Pierson, E., Koh, P. W., Gerardin, J., Redbird, B., Grusky, D., & Leskovec, J. (2021). Mobility-network models of COVID-19 explain inequities and inform reopening. *Nature*, *589*(7840), 82–87. https://doi.org/10.1038/s41586-020-2923-3

de Montjoye, Y. A., Gambs, S., Blondel, V., Canright, G., de Cordes, N., Deletaille, S., Engø-Monsen, K., Garcia-Herranz, M., Kendall, J., Kerry, C., Krings, G., Letouzé, E., Luengo-Oroz, M., Oliver, N., Rocher, L., Rutherford, A., Smoreda, Z., Steele, J., Wetter, E., … Bengtsson, L. (2018). On the privacy-conscientious use of mobile phone data. In *Scientific Data* (Vol. 5, Issue 1, pp. 1–6). Nature Publishing Groups. https://doi.org/10.1038/sdata.2018.286

geoportal.statistics.gov.uk. (n.d.). Population Weighted Centroids. Retrieved February 10, 2021, from https://geoportal.statistics.gov.uk/datasets/workplace-zones-december-2011-populationweighted-centroids

Kitchin, R. (2013). Big data and human geography: Opportunities, challenges and risks. *Dialogues in Human Geography*, *3*(3), 262–267. https://doi.org/10.1177/2043820613513388

Lansley, G., & Cheshire, J. (2018). Challenges to representing the population from new forms of consumer data. *GeographyCompass*, *12*(7), e12374. https://doi.org/10.1111/gec3.12374 -->

ONS. (n.d.). Number of workplaces and employees in Workplace Zones in London, 2009 to 2015. Retrieved July 7, 2020, from https://www.ons.gov.uk/businessindustryandtrade/business/activitysizeandlocation/adhocs/005995numberofworkplacesandemployeesinworkplacezonesinlondon2015

Sahr, K., White, D., & Kimerling, A. J. (2003). Geodesic Discrete Global Grid Systems. In *Cartography and Geographic Information Science* (Vol. 30, Issue 2).

Singleton, A., Longley, P., & Duckenfield, T. (2017). *LondonWorkplaceZones Classification Technical Report*.

Trasberg, T., & Cheshire, J. (2020). *Towards data-driven human mobility analysis*. http://london.gisruk.org/gisruk2020_proceedings/GISRUK2020_paper_39.pdf

Zhou, Y., Xu, R., Hu, D., Yue, Y., Li, Q., & Xia, J. (2020). Effects of human mobility restrictions on the spread of COVID-19 in Shenzhen, China: a modelling study using mobile phone data. *The Lancet Digital Health*, *2*(8), e417–e424. https://doi.org/10.1016/S2589-7500(20)30165-5

Biographies

Louise Sieg is a PhD Student in Geography at UCL. Her research explores the value and provenance of new forms of data applied to geodemographics.

James Cheshire is Professor of Geographic Information and Cartography at UCL, Director of the UCL Q-Step Centre and Deputy Director of the ESRC Consumer Data Research Centre.