Regionalising mobile phone data: attributing time components to optimised regions to create region classifications

Louise Sieg*, James Cheshire†

Department of Geography, University College London

Summary

This paper outlines the development of a classification for a new set of optimised spatial units for the aggregation of mobile phone in-app data. We create regions which aim to strike balance between disclosure control and data granularity, and attribute to each region a time classification reflecting the region's dominant time of mobile phone activity. We aim to provide a methodology for the making of bespoke geographical regions and classifications for the use and dissemination of sensitive granular data in research.

KEYWORDS: Regionalisation, Geospatial 'Big Data', Urban Analytics, Mobile In-app Data

1. Introduction

Data gathered from mobile phones informs research on topics ranging from retail behaviour to epidemiological modelling (Shepherd et al., 2021; Deville et al., 2014). However, multiple studies have demonstrated the need for spatial and temporal aggregation to prevent the disclosive use of such data (De Montjoye et al., 2018; Zhang and Bolot, 2011). Quantifying the impact of this disclosure control on research outputs remains challenging since data aggregated for dissemination purposes tends to lose significant granularity and be subjected to the effect of the modifiable areal unit problem (MAUP) (Openshaw, 1981).

To reduce information loss when data is repurposed for research, optimised geographical units can be created to best fit the data (Kishore et al., 2020). This is the case with the UK Census Output Areas (OA), which were constructed to closely represent the underlying census populations (Martin, 2010). However, no such propositions have yet been made for the regionalisation of new forms of data, such as mobile data.

This paper seeks to demonstrate the utility of bespoke regions that have been constructed to fit underlying terrain and best represent data distribution, whilst remaining non-disclosive by grouping low counts into larger areas. We seek to create region classifications for analysis by first focusing on temporal characteristics. We first introduce the region-making methodology before conducting exploratory analysis to support the time-based region classification.

2. Data

The data comprises GPS location data and timestamps collected by mobile apps. The data is stored securely in an ISO27001 facility with access restricted to researchers who have undertaken appropriate training. The analysis is conducted using an aggregated data product, removing all identifiers: we create an "activity count" consisting of the count of unique devices per region. This paper examines Greater London, with a focus case study on Camden borough. The timeframe is September 2019.

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

3. Methodology

1. Data Regionalisation and Aggregation

To mitigate the effect of aggregation on data granularity and quality, we develop data driven bespoke regions and propose a new methodology. Using the in-app data, we generate regions built upon the Uber H3 hexagonal indexing system (Uber Technologies, 2018). We first aggregate the original datapoints to H3 hexagons (hex) of resolution 10 (65.9 metre edge and 15,047 m2 area) and assign land use and administrative characteristics to each hex. We then filter all hexes with potentially disclosive activity counts (<10) and assign to each low count hex the neighbour which most closely matches its characteristics. We retrieve all hexes that share a connection, treating all hexes as nodes and connections as edges, which returns list of hexagons belonging in similar groups (of similar land use, low activity, and administrative membership). These groups are thus merged into higher level regions, summing their respective activity counts to prevent disclosure, and respecting underlying geographic characteristics.

The resulting H3-based region areas average 62,171 m^2, a comparable scale to the OA (62,500 m^2) but befitting our in-app dataset rather than census statistics. By aggregating a day of data to these H3-regions, OA, and tiles of similar scales (OSGB 250mx250m), and comparing counts, we find that aggregating our data points to H3-regions preserves 25% more data than aggregating them to OA, highlighting the benefits of optimised zoning for aggregation.

	Data omitted due to low counts (%)	Spatial units omitted (%)	**Remaining data count**
OSGB250	8%	42%	**474,499**
OA	8%	51%	**510,849***
H3-based regions	3%	27%	**665,700**

**25% less than H3-based regions remaining data count.*

Figure 1: Summary of comparison between aggregation methods. The H3-based regions preserve more areas and data points than its counterparts.

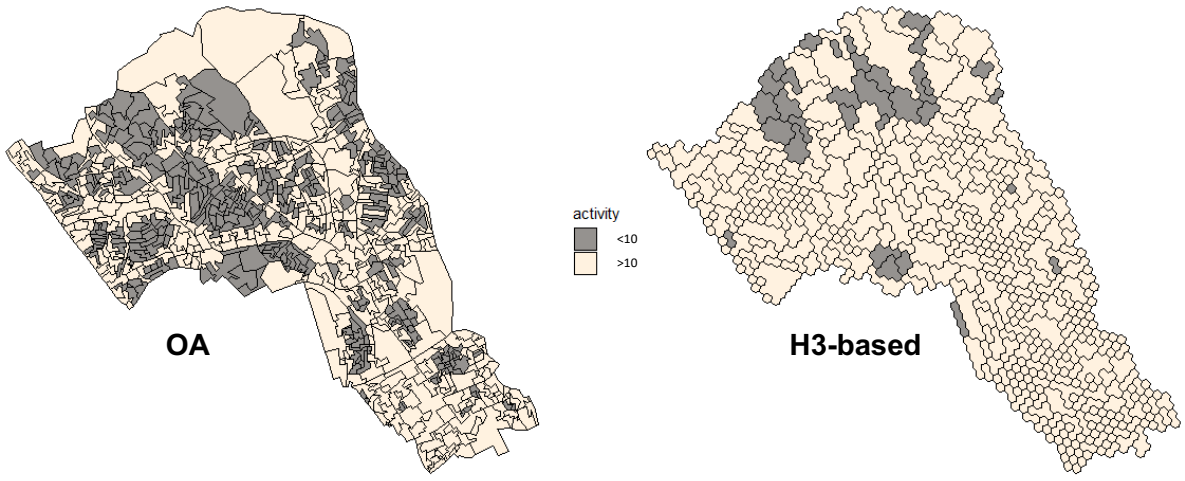

Figure 2: Comparison of areas omitted due to low counts (grey) for OA and H3-based aggregated for the London Borough of Camden

2. Exploratory analysis

Understanding the data's distribution over time can give an indication of each region's patterns and

contribution to the overall activity and help create region classifications to support analysis. We use the timestamp element of our dataset to obtain each region's most active hour of the day and most active day of the week. We average activity counts per region over the course of September 2019 to obtain stable measures of activity per hour and day per region. To obtain the most active hour per region, all event's timestamps are rounded to the hour, and the most active hour per region (over the course of the month) is kept as the region's "top hour". The same thing is done per day of the week, after creating a "weekday" variable from the timestamp's date.

3. Creation of a time-based classification

From the exploratory analysis' results, a time classification is attributed to each region based on its most active times. We present the results of this exploration at the London scale and highlight its functionality by mapping the borough of Camden. The classification is representative of the dataset used for the making of the regions, as it uses this data's activity patterns as its main component.

4. Results

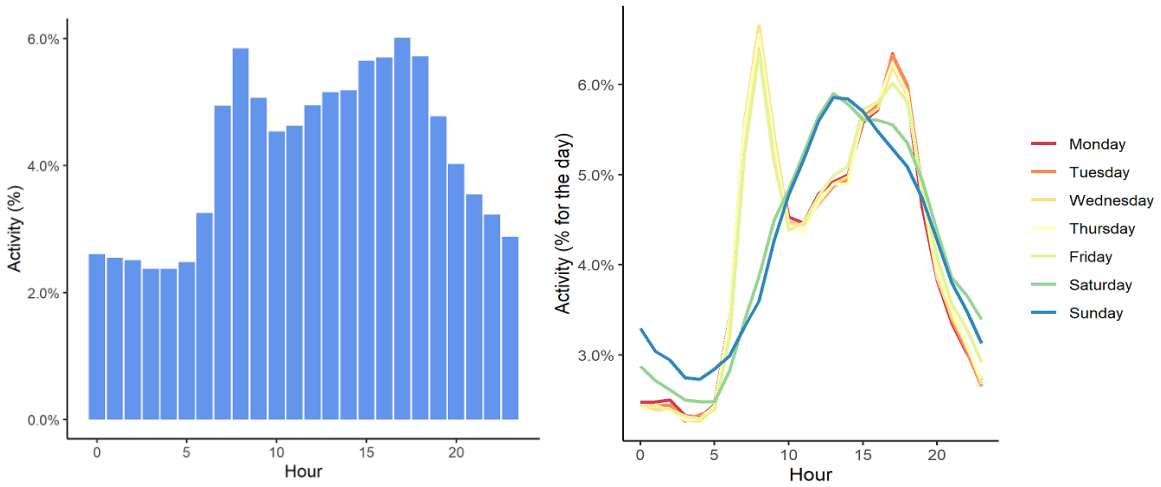

Figure 3 Proportions of hourly activity of a typical day in London (left) and hourly of activity for a typical week (right).

Our exploratory analysis highlighted two distinctive daily patterns: the weekday pattern (from Monday to Friday), which displays two main peaks of activity around rush hour, and the weekend pattern, which is unimodal and increased until midday. These patterns, though expected, help illustrate our data's behaviour on the London scale and inform how to decompose our time-attributes. Considering **Figure 3**, we see that activity is driven differently whether on a weekday or weekend – geographically, distinguishing regions of weekend or weekday dominance can help characterise the type of activity they contain.

Furthermore, night activity (6pm-5am), though less dominant than day-time activity (6am-5pm), represents 39% of activities across the study period. As night-time mobility and activity are less studied than daytime, the use of new granular datasets and classifications could help distinguish areas particularly driven by the nighttime economy (NTE) (Kim, 2020). We thus distinguish regions of nighttime or daytime dominance.

Figure 4 represents the H3-regions at the Camden scale, coloured by resulting assigned time attributes. The time attributes summarize a region's most active time, combining weekday attribute and daytime attributes. We see that most of the green space (Hamstead Heath, Primrose Hill, Regent's Park) activity occurs over the weekend, during the day, whereas Soho is a weekend night hub. The Bloomsbury and Fitzrovia areas, mostly study and work neighborhoods, are dominated by weekday

activity, and Camden high street (above primrose hill) is most active at night. This level of granularity, offered by the in-app dataset and bespoke aggregation, allows for better illustration of smaller area activity patterns.

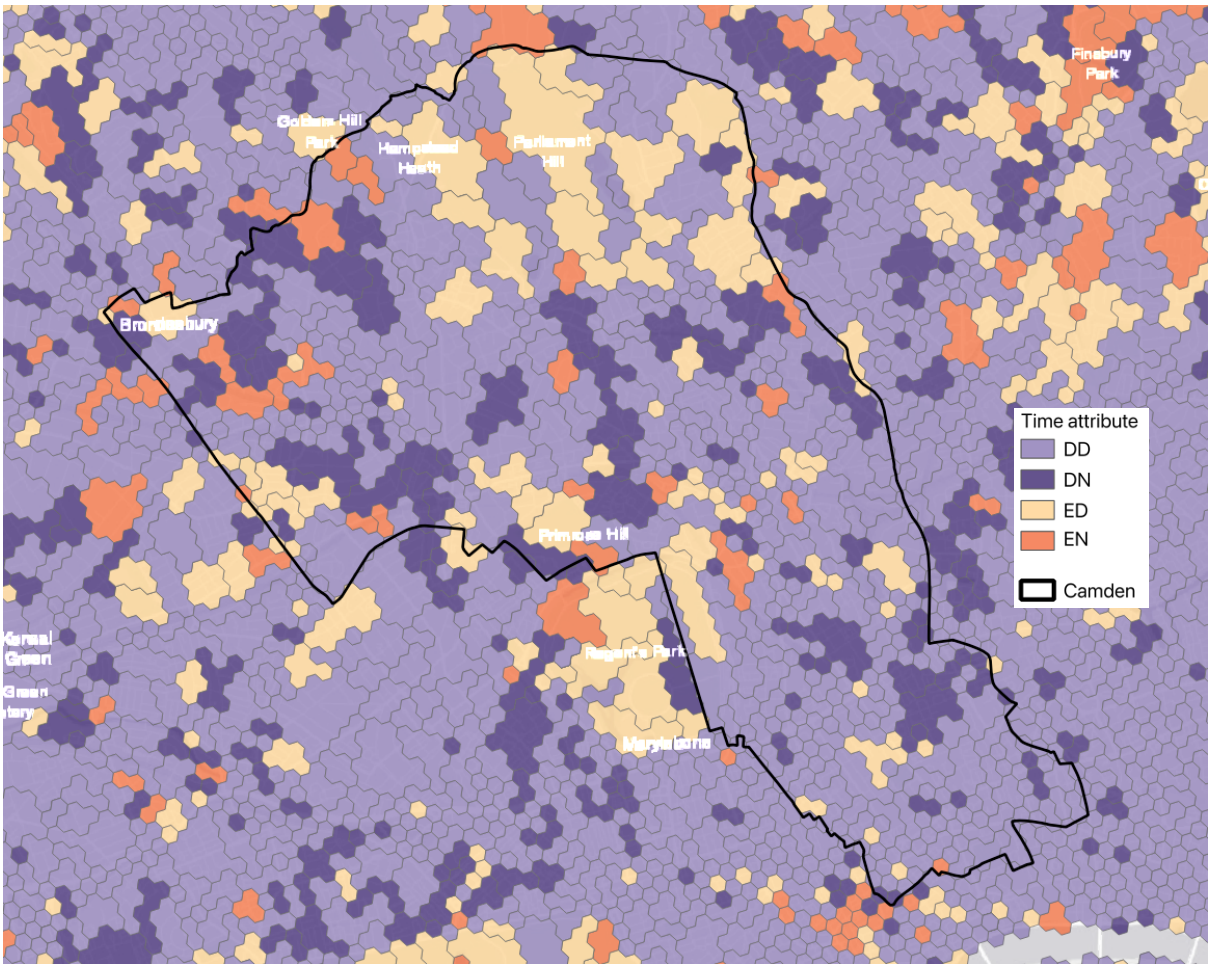

Figure 4: Camden borough (outlined) coloured by time attributes: Weekday-Daytime (DD), Weekday-Nighttime (DN), Weekend-Daytime (ED) and Weekend-Nighttime (EN)

5. Discussion

This paper outlines a methodology for the making of optimised spatial units for data aggregation. A time attribute was created to introduce an initial classification to categorise the outputted regions. This classification adds another layer of information to the bespoke regions. These regions being data-specific, they preserve in-app data granularity compared to other aggregation methods. However, spatial units should also be optimised for research objective (Kishore, 2020). This study aims to demonstrate how the time attribute can help inform on which further specifications could be made for better use-case regionalisation. For instance, new regions could be generated with a focus on nighttime activity, which would be less granular but more appropriate for aggregating low night counts in places of interest.

Moving forward, we aim to further develop our region classification for the characterisation of our data, by combining, clustering and ranking the time-attribute with other information (such as points of interests or land use). We also aim to further develop and package the regionalisation methodology for application to other sensitive point datasets.

6. Acknowledgements

This research was funded by the ESRC. We thank the UCL Research Ethics Committee for reviewing the project.

References

Deville, P. *et al.* (2014) 'Dynamic population mapping using mobile phone data', 111(45), pp. 15888–15893. Available at: https://doi.org/10.1073/pnas.1408439111.w

Kim, Y.-L. (2020) 'Data-driven approach to characterize urban vitality: how spatiotemporal context dynamically defines Seoul's nighttime', *International Journal of Geographical Information Science*, 34(6), pp. 1235–1256. Available at: https://doi.org/10.1080/13658816.2019.1694680.

Kishore, N. *et al.* (2020) 'Measuring mobility to monitor travel and physical distancing interventions: a common framework for mobile phone data analysis', *The Lancet Digital Health*, 2(11), pp. e622–e628. Available at: https://doi.org/10.1016/S2589-7500(20)30193-X.

Martin, D. (2010) 'Optimizing census geography: the separation of collection and output geographies', *http://dx.doi.org/10.1080/136588198241590*, 12(7), pp. 673–685. Available at: https://doi.org/10.1080/136588198241590.

de Montjoye, Y.A. *et al.* (2018) 'On the privacy-conscientious use of mobile phone data', *Scientific Data*, 5(1), pp. 1–6. Available at: https://doi.org/10.1038/sdata.2018.286.

Openshaw, S. (1981) 'The modifiable areal unit problem', *Quantitative geography: a British view*, 68 A(2), pp. 60–69. Available at: https://doi.org/10.4157/GRJ1984A.68.2_71.

Shepherd, H.E.R. *et al.* (2021) 'Domestic and international mobility trends in the United Kingdom during the COVID-19 pandemic: an analysis of Facebook data', *International Journal of Health Geographics*, 20(1), p. 46. Available at: https://doi.org/10.1186/s12942-021-00299-5.

Uber Technologies Inc. (2018) *H3: Uber's Hexagonal Hierarchical Spatial Index | Uber Blog*. Available at: https://www.uber.com/en-GB/blog/h3/ (Accessed: 28 November 2022).

Zang, H. and Bolot, J. (2011) 'Anonymization of location data does not work: a large-scale measurement study', *Proceedings of the 17th annual international conference on Mobile computing and networking - MobiCom '11*, p. 145. Available at: https://doi.org/10.1145/2030613.2030630.

Biographies

Louise Sieg is a PhD Student in Geography at UCL. Her research explores new strategies of data regionalisation for dissemination.

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