The use of in-app data to drive geodemographic classification of activity patterns

Mikaella Mavrogeni*[1], Paul Longley†[1],
Justin van Dijk‡[1]

[1]Department of Geography University College London

GISRUK 2023

Summary

We use location data from multiple mobile phone applications to describe daily, weekly, seasonal and annual activity patterns. Geodemographics, or 'the analysis of people by where they live', provides an organising framework, extended to represent the ways in which neighbourhood residents interact with workplaces, recreational and leisure destinations and transport infrastructure. We evaluate how in-app location data can be incorporated into geodemographic analysis to better understand the flux of activity patterns that characterise densely populated areas throughout the day. Limitations and net benefits of in-app location data are critically assessed to evaluate the ways in which activity-based geodemographics are robust, effective and safe to use when characterising the population at large.

KEYWORDS: geodemographics, big data, temporal analytics, in-app data, geospatial

1. Introduction

Geodemographics present a conventional organising framework for representing the ways in which neighbourhoods are differentiated. They use a range of techniques for summarising large volumes of data into summary profiles that policymakers find helpful in making resource allocation decisions (see **Figure 1** for an illustrative geodemographic classification).

* mikaella.mavrogeni.19@ucl.ac.uk
† p.longley@ucl.ac.uk
‡ j.t.vandijk@ucl.ac.uk

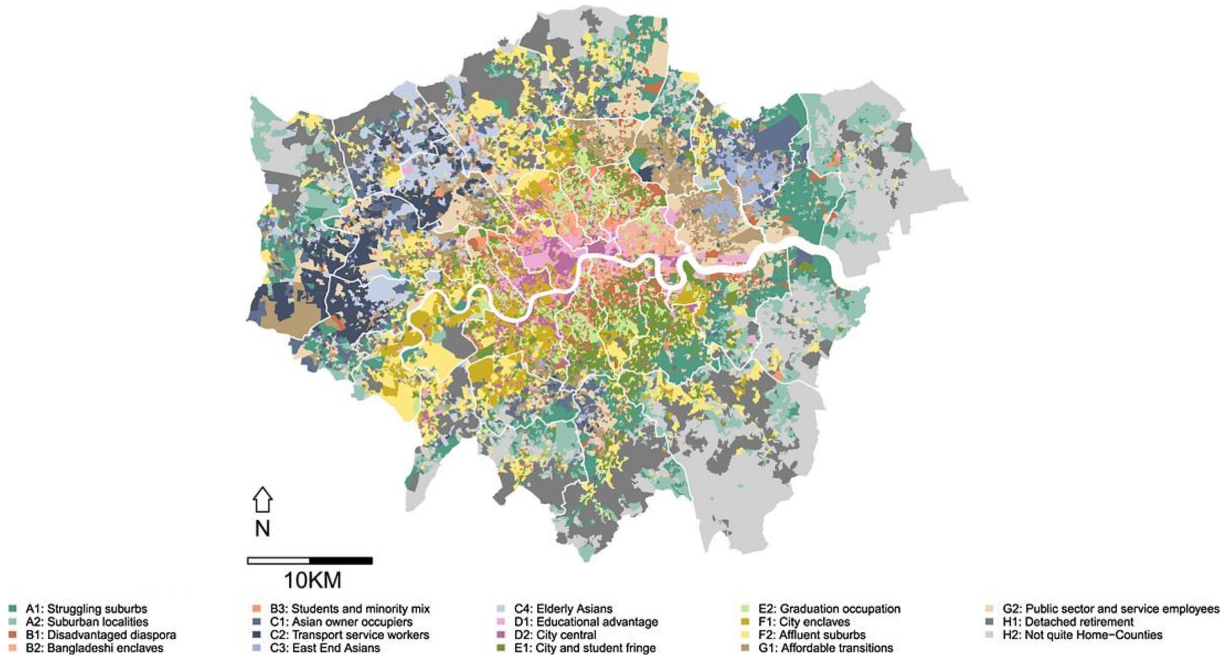

Figure 1: London Output Area Classification (Singleton and Longley, 2015)

There has been long-standing research of geodemographics in the UK with popular geodemographic classifications including the Output Area Classification and the Workplace Zone Classification (Vickers and Rees, 2007; Gale et al., 2016; Cockings et al., 2020). These classifications use variables pertaining to individuals at their normal places of residence or work, albeit not identified for specific points in time – Output Area Classification at place of residence and Workplace Zone Classification at place of employment. They are therefore not fully representative of the full activity patterns of the population beyond nighttime residence and work. This paper aims to evaluate time-stamped in-app location data to address the research gap introduced above. By developing a classification that combines people's place of residence, work, and any intermediate activity patterns they engage in, geodemographic classifications can uncover the differences in temporal profiles of activities. This paper therefore firstly presents the concept of space-time convergence to explain population mobility and highly variegated lifestyles. Secondly, the emergence of big data, and specifically mobile phone location data is presented as an opportunity to capture a fuller range of everyday activities of app users. Thirdly, the benefits of location data are presented against their drawbacks to then evaluate what constitutes a reliable, valid and ethical research using such data. Combined with the argument for a more spatially and temporally granular geodemographic classification, this paper demonstrates that in-app location data's potential of advancing geodemographics is highest when combined with conventional data sources.

2. Time Geographies; Space-time convergence

To fully understand the individual, not only different activities must be considered together, but there is also a need to understand that location has both space and time components. Torsten Hagerstrand (1970) was credited in thinking about movement in space and time by emphasising the importance of accounting individuals/households at a micro-level rather than at a large aggregate scale when it comes to recording changes in their movements. Hagerstrand's (1970) argument is that time needs to be taken into account along with space, something that is clearly demonstrated in **Figure 2**.

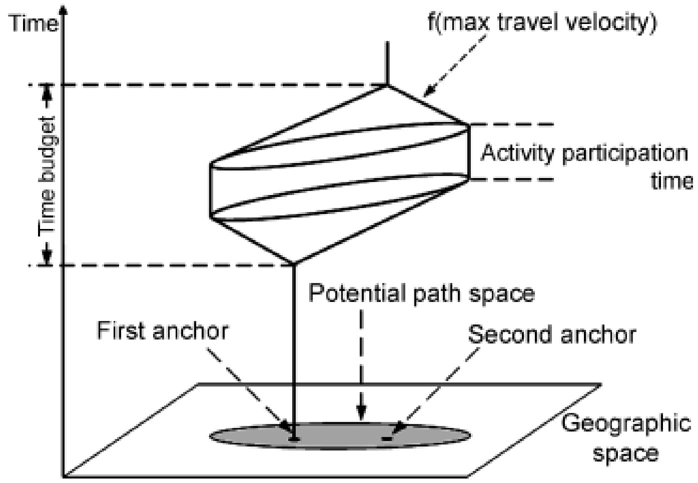

Figure 2: Simple model demonstrating the space-time concept developed by Hagerstrand (Miller 2008)

The changing nature of place was a result of occurrences like transportation space-time convergence with the invention of airplanes, high mobility and telepresence. Despite people sharing workplaces and homeplaces, their lifestyles can be dramatically different due to mobility and space-time convergence i.e., fragmentation of activities (Miller 2007). It therefore becomes crucial to take space-time convergence into consideration when analysing the population's mobility patterns. A facilitator in achieving that is big data, and more specifically location data that will be discussed in the following sections.

3. Big data; a new opportunity for GIS

'Big data' is data that is huge in volume, diverse in variety and high in velocity, and has influenced the way in which data is managed (Kitchin, 2013). Big data sources include location data that is automatically collected from smart-phone devices which can help track population movement (Deville et al., 2014). As a tool that is integrated in people's lives, especially in developed countries, the smartphone allows for near real-time tracking of activities which can be used to augment current data gathering methods (Raento Google et al., 2009).

3.1 Benefits of mobile phone location data

Having access to location data facilitates the understanding of individuals' activity patterns near real-time and allows researchers to characterise the population based on lifestyle and day-to-day activities. This in turn solves the problem of obsolescence and leads to greater temporal granularity with regards to the periods covered by the available data. There is no longer the need to wait for census or survey datasets which are collected at specific time periods. Rather than just focusing on data from one point in time in the past, 'big data' allows analyst investigation of where people go, what activities they are involved with and what other areas they interact with near real-time. Population density is constantly changing throughout the day, with the night-time population significantly differing from the day-time population.

3.2 Drawbacks of mobile phone location data

Because such data are produced for marketing purposes, quality control that complies with academic standards is not a requirement and thus there are certain challenges involved when repurposing the data (Lansley and Cheshire, 2018). Firstly, data are not random but instead self-selective depending on who owns a smart phone and uses any of the range of apps that can be installed on it, which then becomes an issue of representation. Secondly, the large volumes of data produced might lead to technical difficulties when storing and managing the information, which might act as a barrier to researchers utilising the data. Thirdly, the GPS location provided by the smartphone might be subject to accuracy concerns (Raento Google et al., 2009; Wu and Zhao, 2015). Lastly, there are ethical concerns associated with location data because it consists of personal data that can be disclosive when released in the form of timed location with co-variates (Georgiadou et al. 2019). As de Montjoye et al. (2013:1) emphasised, four spatio-temporal points can be enough to uniquely identify 95% of individuals.

4. Empirical analysis: location data and Geodemographic classification

Locationally time-stamped in-app impressions can reveal activity patterns which when combined with the place of work and residence can facilitate the creation of a geodemographic classification that is representative of the changing dynamics of today's population. In addition to helping reach greater temporal granularity, the individual-level location data will facilitate greater spatial granularity. The most granular unit at which geodemographic classifications are produced is the postcode level in the private sector, and the output area level in the public sector. Utilising data provided at the most granular level possible, increases the potential of moving away from the one-size-fits-all classification and overgeneralisations, towards location-specific characterisations.

Completely shifting away from conventional data to dynamic data shows a misunderstanding of what the two data types have to offer and thus both should be used together in recognition of their distinct benefits (Kitchin, 2013). Additionally, research outputs that aggregate the results having the GPS accuracy threshold in mind, will not suffer as much from the limitations associated with GPS technologies. With regards to the biggest drawback, ethical concerns can be mitigated in relation to the processing of the data, the output of research results, and researcher training by accessing the data in secure research environments and performing statistical disclosure control procedures prior to releasing any analysis results.

Despite all the drawbacks of location data, there is a need for research to move away from just relying on conventional data since all the population insights are hidden in data collected from our everyday realities. By weighting the benefits against the limitations, location data is still recognised extremely valuable and should be incorporated in geodemographics to characterise the highly dynamic population near real-time.

Our empirical analysis of Greater London uses in-app location data and geodemographic classifications to characterise user probable place of residence, work and leisure activities. This is achieved using the London Output Area Classification (LOAC) and London Workplace Zone Classification (LWZC) to classify sequences of user impressions and develop summary profiles of user origin-destination pairings. Location data also facilitate the creation of temporal profiles that leverage greater understanding of how individuals interact with different spaces at different times of the day and week. **Figure 3** presents the workflow that we develop for the analysis. We present the outline findings in terms of origin-destination and activity pattern matrices for the pre-pandemic year of 2019.

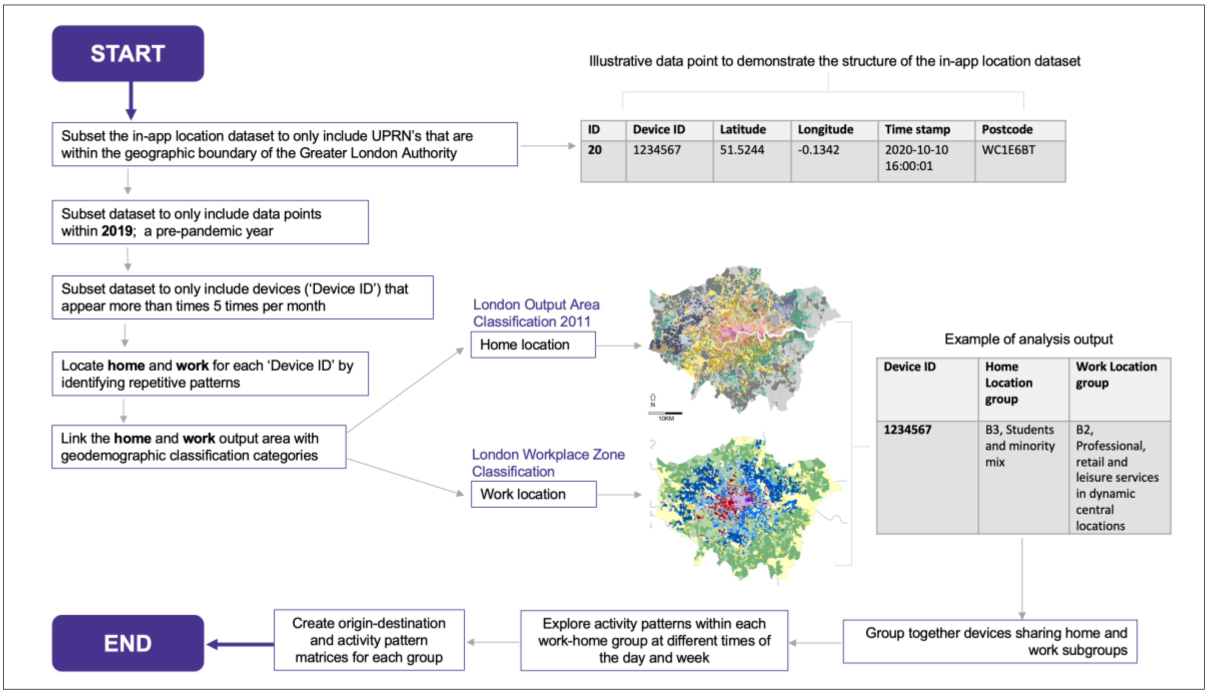

Figure 3: Workflow of analysis used to link location data to London residential and workplace geodemographic classifications

5. Conclusion

This geographically extensive temporal analysis demonstrates the importance of incorporating location data in geodemographic classifications to facilitate the characterisation of the population near real-time based on their everyday activities instead of only their socio-economic characteristics. This can be achieved by creating a series of temporal profiles that represent the population dynamics at different times of the day, week and year. In recognition of the benefits location data can offer, this paper argues in favour of developing new facets of geodemographic research. However, given the several drawbacks and ethical concerns that location data pose, it's crucial for researchers to be cautious when handling and analysing the data. Not only the data might be under-representative of certain population segments, but temporal coverage might not be persistent across users either. Dealing with a data sample instead of census increases the complexity of using location data for geodemographics. Researchers therefore need to perform all necessary steps to ensure that the data is as representative as possible but also combine it with conventional data which provide more complete information of the population's characteristics. Once the discussed factors are taken into consideration, it is feasible to utilise location data to produce near real-time geodemographics of the population.

Acknowledgements

This research is co-funded by **UBEL DTP and Didobi Ltd. Didobi lead Visiting** Professor Matthew Hopkinson has shared industry advice and insight on the application domain developed in this paper.

References

Cockings S, Martin D and Harfoot A (2020) Developing a National Geodemographic Classification of Workplace Zones. *Appl. Spatial Analysis*, 13, 959-983.

de Montjoye YA, Hidalgo C A, Verleysen M and Blondel V D (2013) Unique in the Crowd: The privacy bounds of human mobility. *Sci Rep*, 3, 1376.

Deville P, Linard C, Martin S, Gilbert M, Stevens F R, Gaughan A E, Blondel V D, and Tatem A J (2014). *Dynamic population mapping using mobile phone data*. 111(45), 15888-15893.

Gale C G, Singleton A D, Bates A G and Longley P A (2016) Creating the 2011 area classification for output areas (2011 OAC). *Journal of Spatial Information Science*, 1-27.

Georgiadou Y, By, R A de, and Kounadi O (2019). Location Privacy in the Wake of the GDPR. *ISPRS International Journal of Geo-Information 2019*, 8(3), 157.

Hägerstrand T (1970) What about people in Regional Science?. *Papers of the Regional Science Association* 24, 6–21.

Kitchin R (2013) Big data and human geography: Opportunities, challenges and risks. *Dialogues in Human Geography*, 3(3), 262-267.

Lansley G and Cheshire J (2018) Challenges to representing the population from new forms of consumer data. *Geography Compass*, 12(7), e12374.

Miller H (2007) Place-based versus people-based Geographic Information Science. *Geography Compass*, 1(3), 503–35.

Miller H (2008) Time Geography. In: Shekhar S, Xiong H (eds) *Encyclopedia of GIS. Springer*, Boston, MA.

Raento Google M U, Antti Oulasvirta L and Eagle N (2009). Smartphones An Emerging Tool for Social Scientists. *Sociological Methods & Research*, 37, 426-454.

Singleton A D and Longley P (2015) The internal structure of Greater London: a comparison of national and regional geodemographic models. *Geography and Environment*, 2, 69-87.

Vickers D, Rees P (2007) Creating the UK National Statistics 2001 output area classification. *Journal of Royal Statistical Society A*, 170, 379-403.

Wu C and Zhao Y (2015) Human Mobility Enhances Global Positioning Accuracy for Mobile Phone Localization. *IEEE Transactions on Parallel and Distributed Systems*, 26(1).

Biography

Mikaella Mavrogeni is a PhD student co-funded by UBEL DTP and Didobi, with project title ‘Real-time Geodemographics for business and service planning’. Research interests include using geo-spatial data to analyse population mobility such as deriving activity patterns from location data and creating origin-destination matrices from public transport data.

Paul Longley is Professor of Geographic Information Science at University College London, where he directs the Economic and Social Research Council-funded Consumer Data Research Centre.

Justin van Dijk is a Lecturer in Social and Geographic Data Science in the Department of Geography at University College London. His primary research interests are grouped around the analysis and visualisation of large-scale spatial data and socio-spatial inequalities.