Deriving flow patterns from GPS in-app mobile phone data

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

[1]Department of Geography University College London

GISRUK 2024

Summary

GPS location data can reveal information about individuals' everyday lives, something that conventional data sources like census data cannot do. However, one major limitation of GPS location data is that almost always the location will be recorded with a level of error, known as positional uncertainty. This paper works around the above limitation by aggregating the data at the MSOA level and performing origin-destination analysis. Origin-destination matrices are created to investigate interaction flows and reveal insights on MSOA level connections. We discuss how the analysis can benefit policymakers and public transport providers.

KEYWORDS: interactional geographies; temporal analytics; in-app location data; positional uncertainty

1. Introduction

GPS in-app mobile phone data offer several opportunities for investigating the lives of individuals beyond their place of residence or employment. However, GPS inaccuracies may constrain the analysis by introducing problems of positional uncertainty and thus limiting the confidence with which conclusions from disaggregate location data are made. This research project uses data obtained from HUQ, where 85% of location points in the data are recorded with accuracy of less than 100 metres and come from GPS or assisted GPS where GPS triangulates its position between many GPS satellites, whilst the other 15% of location points are recorded with accuracy of more than 100 metres and come from non-GPS locations. This paper demonstrates how triangulated GPS in-app mobile phone location data can be used to benefit the analysis through the utilisation of origin-destination matrices (OD-matrices) at the MSOA level that can work around the limitations of positional uncertainty. Origin-destination flows are recognised as an extremely important area of research for inferring trip purpose and informing transport planning and policy decisions (Bachir et al., 2019; Iqbal et al., 2014; Ge and Fukuda, 2016). This paper therefore firstly presents the constraints associated with positional uncertainty of GPS data and then mentions the reasons behind the choice of MSOA level analysis as well as OD-matrices. This is followed by the methodology for extracting origin and destination locations from a GPS mobile phone dataset and the analysis of OD-matrix maps at the MSOA level. Lastly, the paper explores several applications including hourly interaction flows as well as location-specific flows, followed by conclusions about the research implications and contributions to the field of geospatial analytics.

2. Limitations of Positional Uncertainty

There is almost always a percentage of error due to GPS device inaccuracies (Ranacher and Brunauer 2015; Djuknic and Richton 2001). Most location points are recorded with a GPS accuracy between 0

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

and 100 metres, with a few location points having inaccuracy of more than 100 metres, depending on factors such as number of satellites in sight and urban canyoning (Kumar and Dutt 2020; He et al. 2017). GPS accuracy will affect the analysis most when the data are in a disaggregate form where, if systematic, a few metres of GPS inaccuracy may completely shift activity hotspots. At an aggregate scale, the less granular the analysis, the smaller the likely effect of GPS inaccuracies, as the likelihood of a location point falling into another polygon due to its inaccuracy is lower. However, with lower granularity, analysis can only be undertaken at a less detailed level, thus affecting the insights that it can offer to policymakers and analysts. Therefore, the aim should always be to minimise the trade-off between the two to ensure that spatial granularity is high enough to provide valuable insights but low enough to reduce the effect of GPS inaccuracies. This paper undertakes analysis at the MSOA level to diminish the issue of positional uncertainty, and in recognition of the trade-off mentioned above, it carries out analysis of recorded origin-destination flows to maximise the information gain from aggregated location data. MSOAs include 2,000 - 6,000 households and big enough areas to ensure there are enough interaction flows between most MSOA pairs for scientific disclosure control purposes.

3. Method

OD-matrices are a well-established and extensively applied spatial network analysis method, which gained significant traction, especially in the last two decades with the increasing accessibility of mobile phone data (Van Dijk et al., 2021; Vanhoof et al., 2021; Demissie et al., 2019; Vij and Shankari, 2015; Zhong *el al.*, 2014; Calabrese, 2011). The relationships between different origins and destinations (OD-pairs) can be depicted by direct lines, connecting these points using Euclidean distances (see **Figure 1**). A sequential colour palette is used to represent the total number of interaction flows between two MSOAs.

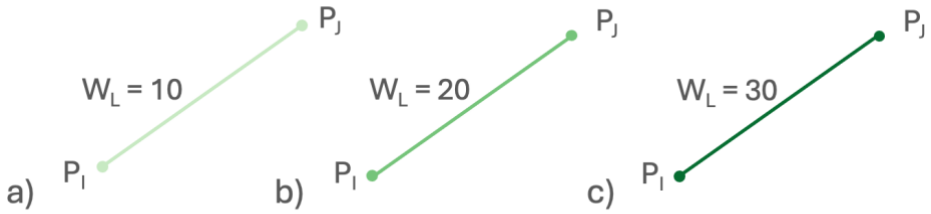


Figure 1: Sequential colour palette used to represent the total number of interactions between MSOAs P_I and P_J is a) 10, b) 20 and c) 30.

The UK-wide in-app mobile phone location dataset from HUQ is filtered down to London for 2019 for this study, because London comprises of 54% of all datapoints and thus it is the best-represented region in the dataset. According to the mid-2019 population estimates from the ONS, London's population was 8.96 million (ONS, 2020), whilst the total number of devices in the HUQ data were just a tiny fraction of that (1.25%). By shifting the analysis to focus on interaction flows between places instead of people, we can avoid some of the representation problems that individual analysis suffers.

The dataset is sorted by device and in chronological order, to facilitate the creation of the OD interaction flows. **Table 1** illustrates the re-structured dataset that links a location (i.e., origin) the next visited location (i.e., destination). The table is created using synthetic data to illustrate how the dataset was manipulated to facilitate interaction flow analysis.

Table 1 Wide format of in-app mobile phone dataset used to derive OD matrices (based on synthetic data).

Device ID	Visit longitude	Visit latitude	Leaving datetime	Next visit longitude	Next visit latitude	Datetime next	Time elapsing (hours)
1234	0.20	51.5	2019-01-08 17:25:00	0.22	51.8	2019-01-08 18:40:00	1.25
1234	0.22	51.8	2019-01-08 22:00:00	0.0567	51.339	2019-01-09 19:25:00	16.42

A threshold of duration less than 2 hours is then set to only consider interactions that were linked within that time frame. The choice of threshold is made to filter out data points that do not have any immediate interactions between locations. **Figure 2** presents the workflow that was followed to obtain the aggregated interaction flow data from the raw GPS in-app location visits.

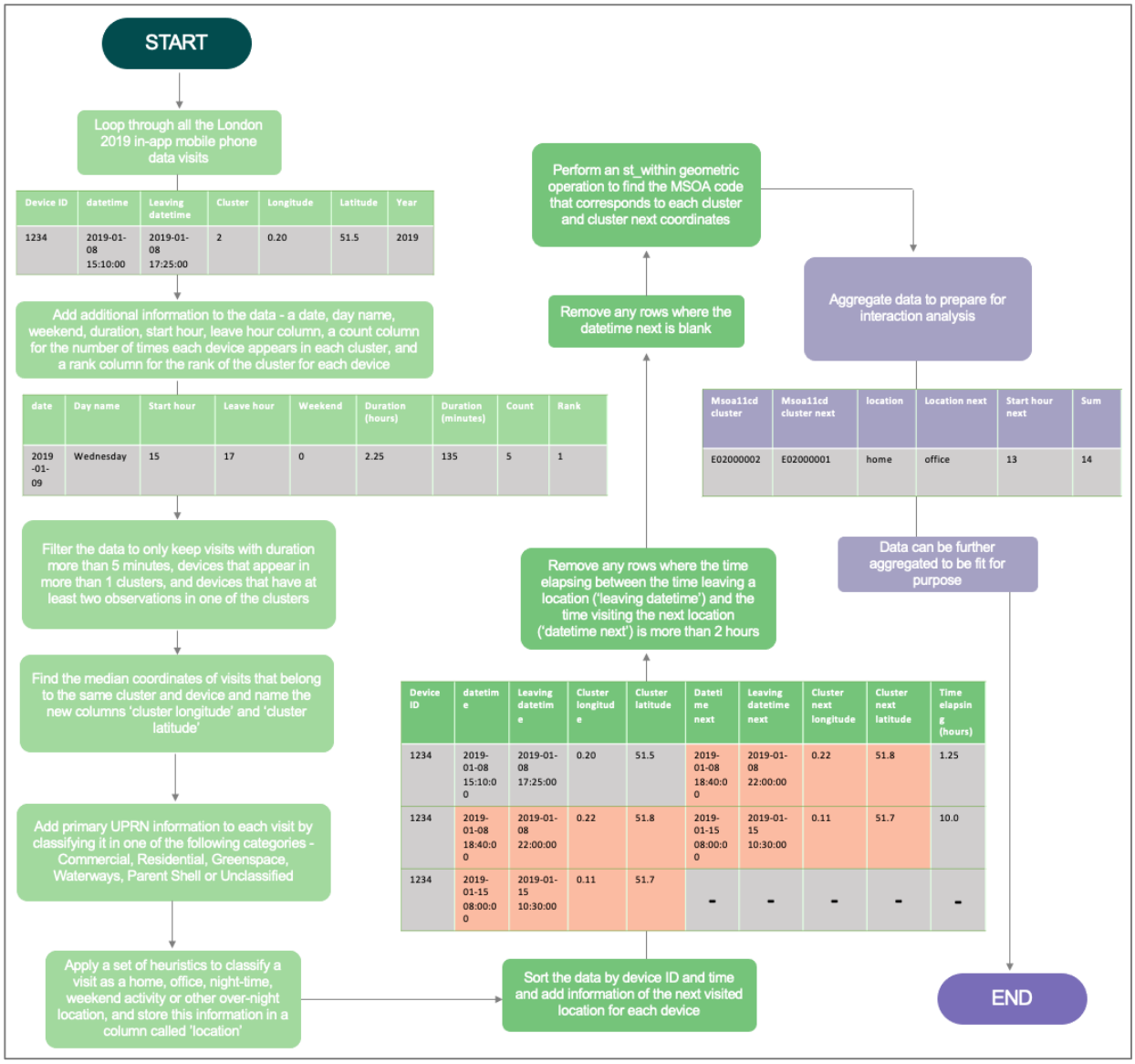


Figure 2: Workflow for obtaining interaction flows from GPS location data visits

Interaction analysis is performed at an MSOA level because MSOAs are formed of big enough areas to ensure that there are several interactions between them, and not much data is lost from disclosure control. To create an OD matrix, it is necessary to first create an aggregated dataset with the total number of interactions between any two MSOAs in 2019 (see **Figure 2**). Aggregated OD-matrices are common in past research that also deals with sensitive data (Van Dijk et al., 2021; Demissie et al., 2019; Ge and Fukuda, 2016). A further breakdown by hour is made to understand how interactions are influenced by the time of day, and lastly, direction was also included to enable the understanding of purpose when combined with time of day. The centroids of the origin MSOA and destination MSOA are linked together using Euclidean distance, which results in a line geometry with a sum of flows attribute. MSOA centroids are used to protect the identity of subjects, and any flows less than 10 are removed for disclosure control purposes.

Interaction analysis displays the most prevalent interaction patterns between the different areas. The overall aim of this analysis is to investigate the areas with the strongest interactions and give meaning behind them from the perspective of place.

4. Empirical Analysis

Figure 3 presents the OD-matrix for the interaction flows between each MSOA in 2019 in Greater London, with darker shades representing a higher number of interactions, and a minimum threshold of 600 used for achieving visual simplicity whilst also ensuring that the most prevalent patterns are displayed (Guo 2009; Gao et al. 2013; Rae 2009).

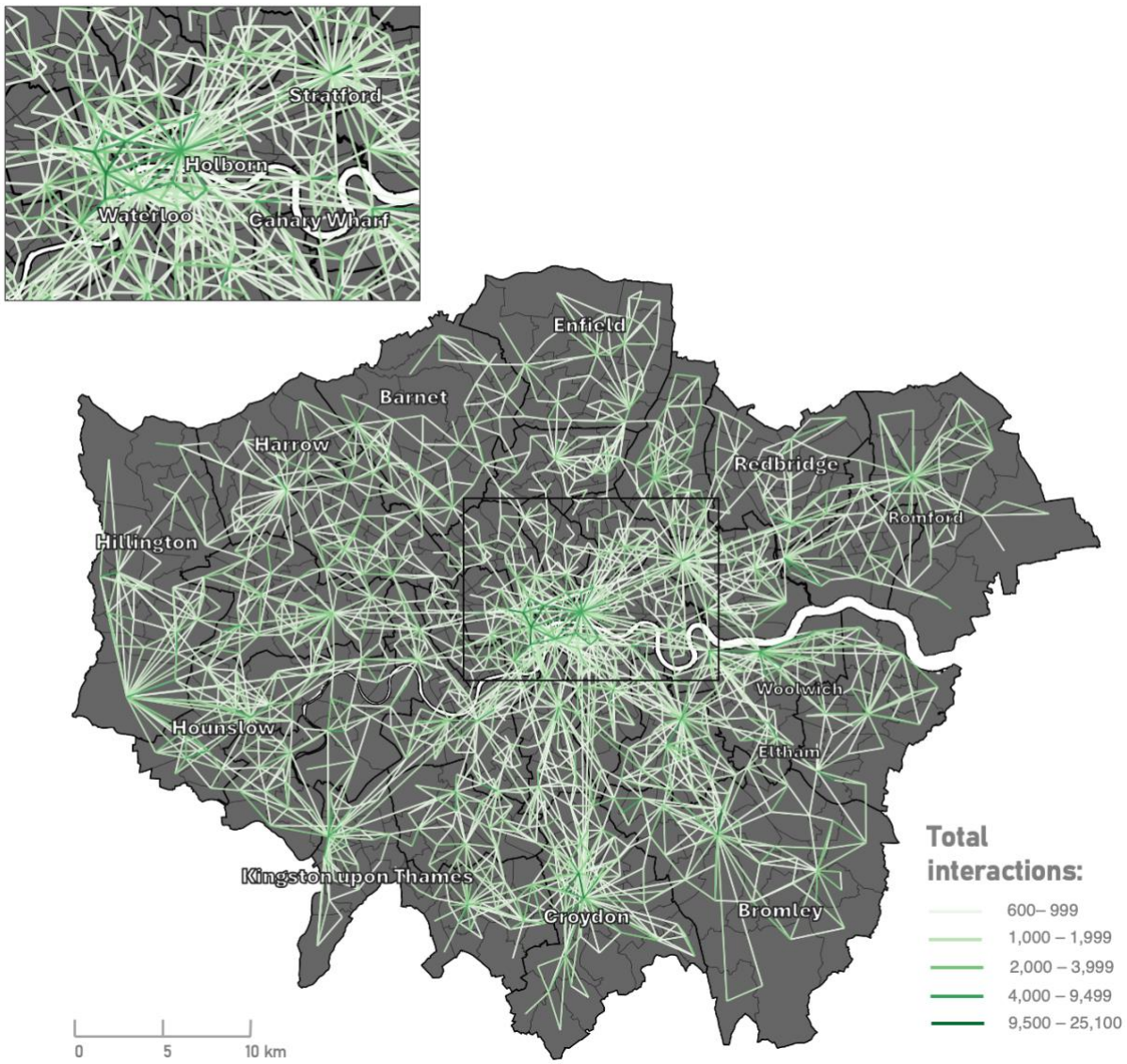


Figure 3 Non-directional origin-destination matrix for the interaction flows between each MSOA in 2019 in London (minimum threshold = 600 interactions).

As seen in the figure, the most connected area is central London, and whilst some regions of greater London are connected to central London, there are many interactions between town centres and nearby MSOAs. To exemplify the aforementioned, Croydon is well connected to central London MSOAs as well as surrounding MSOAs, whilst Eltham in South-East London is mostly connected to a few surrounding areas and Canary Wharf. Whilst some MSOAs might have links with central London MSOAs, these are not visible when only analysing interactions of more than 600 flows. Such patterns can also be explained by travel speed, meaning that some locations might have better transport links and thus better connectivity, making it possible to reach a bigger distance in a given time (2 hours in this case) than locations with weak transport links.

4.1 Interaction flows by hour

Interaction flows can also be analysed by hour or direction in order to get a better understanding on purpose. Breakdowns by hour result in lower interactions between MSOAs, thus the lowest threshold

is set to 60.

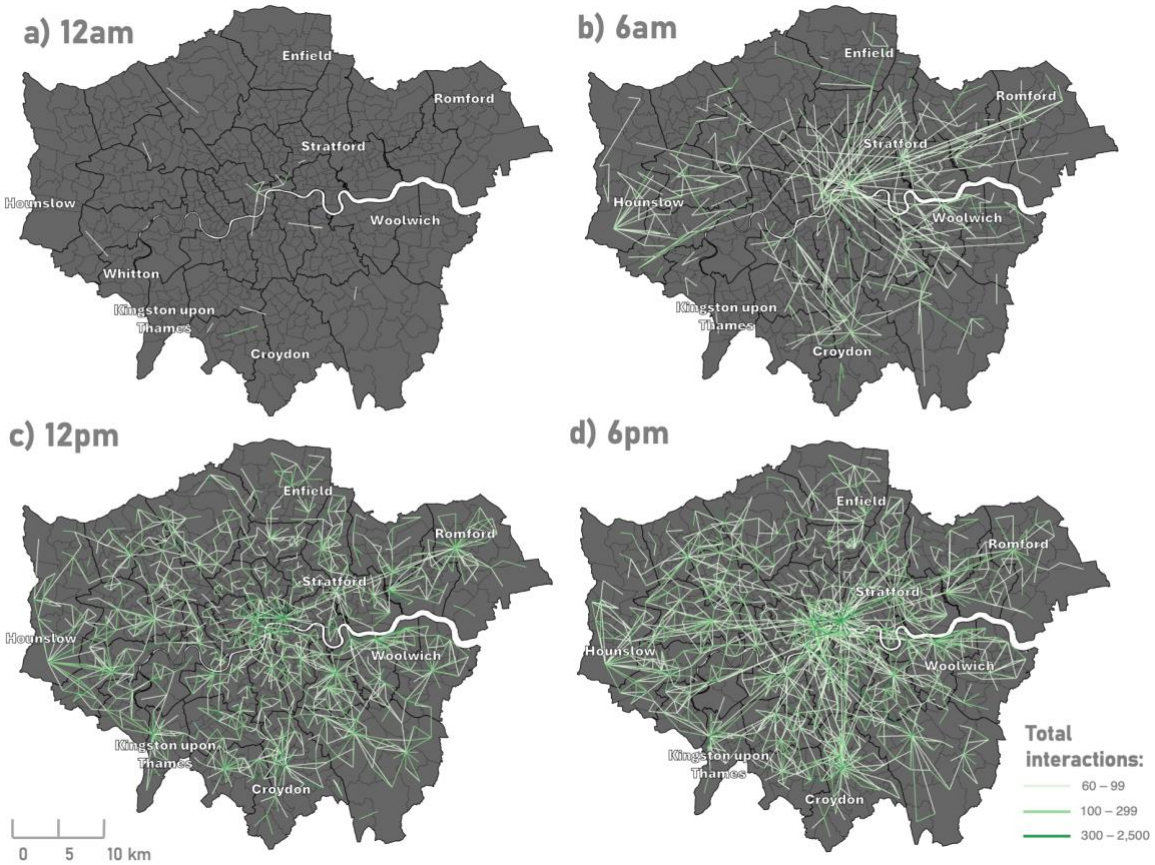


Figure 4 Non-directional origin-destination matrix for the interactions between each MSOA in 2019 in London for a) 12am b) 6am c) 12pm and d) 6pm (minimum threshold = 60 interactions)

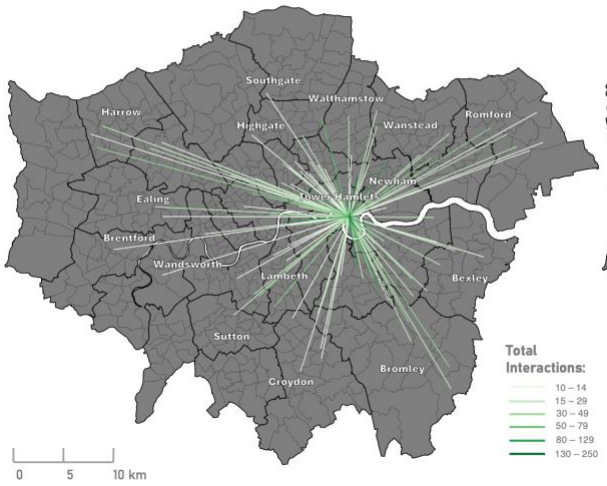
At 12am when interactions are expected to be much lower, the only areas with more than 60 interactions are a) Oxford Street and Covent Garden and b) Heathrow and Whitton (see **Figure 4a**). On the contrary at 6am when interactions are much higher, and commuting to work patterns start to form, there are many prevalent interactions, especially between Suburban MSOAs to central London MSOAs. The main difference between the 6am interactions and 12pm interactions (see **Figures 4b and 4c**)**,** is that 12pm interactions are shorter in displacement as opposed to 6am interactions which are longer in displacement due to their commuting nature. At 6pm, both short and long displacement interactions are visible (see **Figure 4d**), indicating the presence of both commuting patterns and other localised flows.

4.2 MSOA specific analysis; Directional interaction flows by hour for Canary Wharf

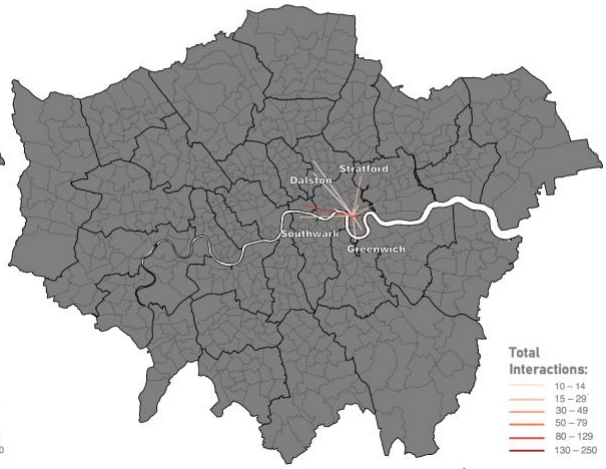
To demonstrate the useability of the above analysis, an MSOA in Canary Wharf is chosen to display the directional interaction flows at four critical time windows: a) 8am, b) 1pm, c) 6pm and d) 11pm. The choice of time is to ensure that a mixture of hours throughout the day are included to assess the function of the area during these hours.

8am

Inbound

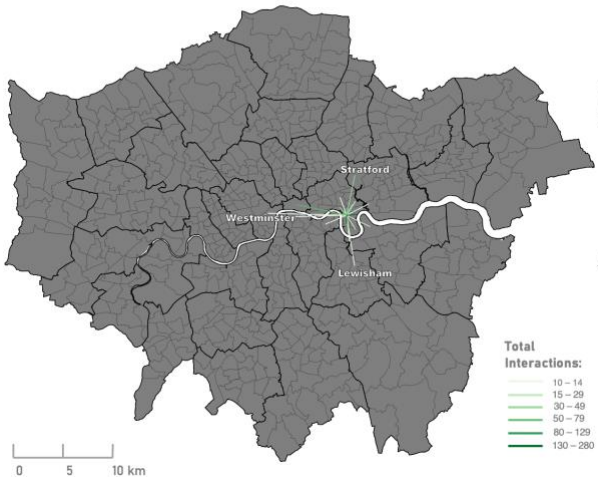


Outbound

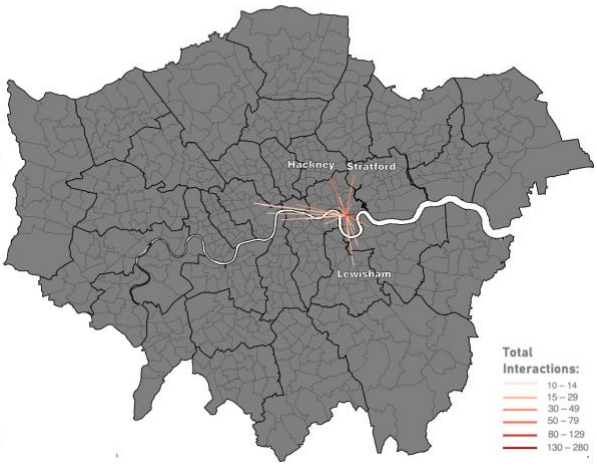


1pm

Inbound

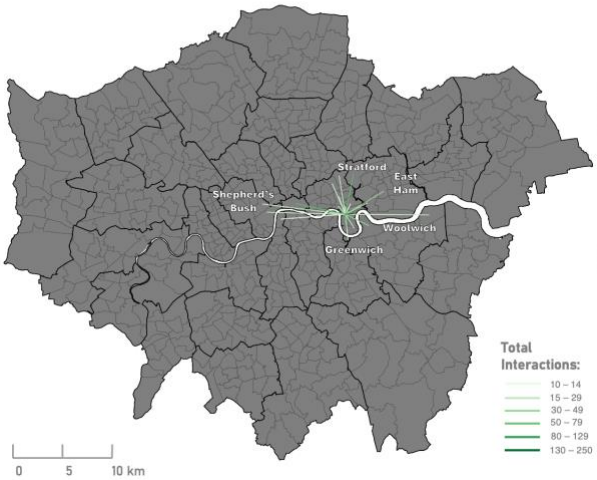


Outbound

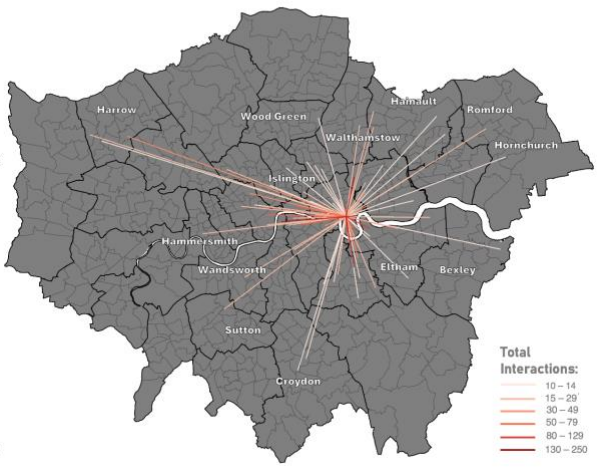


6pm

Inbound



Outbound



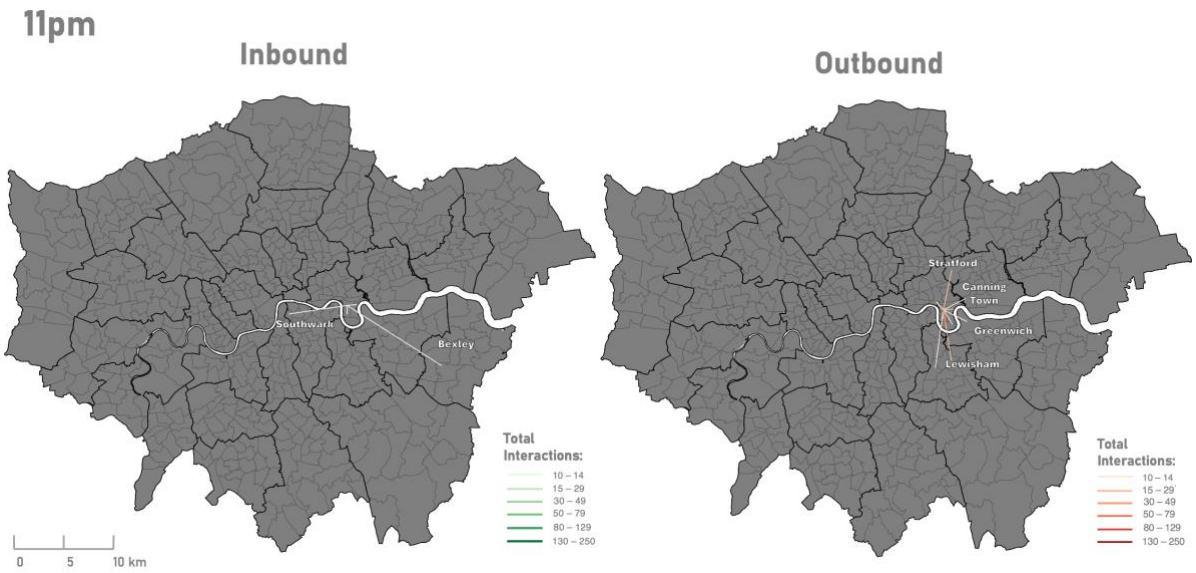


Figure 5: Directional origin-destination matrix maps for the interactions between Canary Wharf and other MSOAs in 2019 in London for a) 8am b) 1pm c) 6pm and d) 11pm

As visible from **Figure 5**, Canary Wharf experiences a big influx of people at 8am, with the main contributions being from Tower Hamlets, Newham and Lambeth. By contrast, at 6pm there is a net outflux of people to several surrounding MSOAs as well as other MSOAS located further away in North-West and South-West London, and to a lesser extent South London. The flows at 1pm and 11pm occur at much smaller volumes than during 8am and 6pm, but are more balanced than the 8am and 6pm flows. Canary Wharf is characterised as a mainly commercial area but also consists of some residential buildings, and whilst at 8am it is expected to experience a net influx of commuters, there are some individuals that might commute from Canary Wharf to other areas thus following the opposite patterns.

Some areas have only a limited range of interactions with surrounding areas, whilst some other areas have a wider range of interactions, connecting with areas further away, depending on the functioning of different labour markets and the integrity of neighbourhood structure. The latter pattern is visible for central city locations such as London Bridge, Westminster and Canary Wharf. Even if the above areas are widely connected, there are noticeable differences between them in terms of the areas they interact with most. This emphasises the benefits of interaction analysis, and a further breakdown by hour facilitates the understanding of when these interactions take place, to then infer purpose.

5. Conclusions

Origin-destination analysis has demonstrated the useability of GPS mobile phone data for understanding connectivity, reach, and links between locations throughout the day, with the potential of focusing on each MSOA individually as illustrated in **Section 4.2**. However, the study of interaction flows doesn't come without its limitations. The problem of GPS inaccuracies still persists even if only data with less than 100 metres accuracy is considered. The aggregation step taken in the analysis helps to partly limit the problem, but even small levels of GPS inaccuracy might mean that the interaction appears to originate/end at a different MSOA than the one it actually originates/ends at. Additionally, the exact route taken is unknown as well as any intermediate visited location. Nonetheless, interaction flow analysis effectively visualises flow patterns and can still offer value for policymakers despite its limitations.

Further research can expand the analysis to explore seasonal variations as well as weekday and weekend interaction flows to provide additional insights for urban planning and transportation management. Additionally, it can focus on enriching the above analysis by incorporating socio-economic and

demographic data from the 2021 census to build a near-real time geodemographic classification that doesn't just focus on understanding place from the socio-economic and demographic composition of its residents, but also from the activity flows that characterise them throughout the day. Interaction flows can be integrated into the formulation of an index of activities deducing the level of connectivity, reach/extent and pupropes for each spatial unit throughout the day. Distance and speed thresholds could also be used to achieve the segmentation of origin-destination flows by mode of transport such as foot, car, bus, train etc. Lastly, this analysis can benefit policymakers and public transport providers by revealing valuable information on the locations and times that public transportation is under or over-provided.

6. Acknowledgements

This research is co-funded by the ESRC **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

Bachir, D., Khodabandelou, G., Gauthier, V., El Yacoubi, M. and Puchinger, J. (2019) Inferring dynamic origin-destination flows by transport mode using mobile phone data. *Transportation Research Part C: Emerging Technologies*, 101, 254–275.

Calabrese, F., Di Lorenzo, G., Liu, L. and Ratti, C. (2011) Estimating Origin-Destination Flows Using Mobile Phone Location Data. *IEEE Pervasive Computing*, 10, 36–44.

Demissie, M.G., Phithakkitnukoon, S., Kattan, L. and Farhan, A. (2019) Understanding Human Mobility Patterns in a Developing Country Using Mobile Phone Data. *Data Science Journal*, 18, 1.

Djuknic, G.M., Richton, R.E. (2001) Geolocation and assisted GPS, *Computer*, 34, 123–125.

Ge, Q., Fukuda, D., (2016) Updating origin–destination matrices with aggregated data of GPS traces. *Transportation Research Part C: Emerging Technologies*, 69, 291–312.

Guo, D. (2009) Flow mapping and multivariate visualization of large spatial interaction data. *IEEE Transactions on Visualization and Computer Graphics*, 15, 6, 1041-1048.

Gao, S., Liu, Y., Wang Y. and Ma, X. (2013) Discovering spatial interaction communities from mobile phone data. *Transactions in GIS*, 17, 3, 463-481.

He, X., Montillet, J.-P., Fernandes, R., Bos, M., Yu, K., Hua, X. and Jiang, W. (2017) Review of current GPS methodologies for producing accurate time series and their error sources. *Journal of Geodynamics*, 106, 12–29.

Iqbal, Md.S., Choudhury, C.F., Wang, P. and González, M.C. (2014) Development of origin–destination matrices using mobile phone call data, *Transportation Research Part C: Emerging Technologies 40*, 63–74.

Office for National Statistics ONS (2020) Population Estimates for the UK, England and Wales, Scotland and Northern Ireland, provisional: mid-2019. Available here [Accessed 01.03.2024].

Rae, A. (2009) From spatial interaction data to spatial interaction information? Geovisualisation and spatial structures of migration from the 2001 UK census. *Computers, Environment and Urban Systems*, 33, 3, 161-178.

Ranacher, P, Brunauer, R, SPEK, S. v. d. and Reich, S. (2015) GPS error and its effects on movement analysis, Instrumentation and Detectors.

Sirish Kumar, P., Srilatha Indira Dutt, V.B.S., (2020) The global positioning system: Popular accuracy measures. *Materials Today: Proceedings*, 33, 4797–4801.

Van Dijk, J.T., Lansley, G. and Longley, P.A., (2021) Using Linked Consumer Registers to Estimate Residential moves in the United Kingdom. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 184, 1452–1474.

Vanhoof et al. (2021) Performance and Sensitivities of Home Detection on Mobile Phone Data in Hill, C.A., Biemer, P.P., Buskirk, T.D., Japec, L., Kirchner, A., Kolenikov, S., Lyberg, L. (Eds.), 2021. Big data meets survey science: a collection of innovative methods. Wiley, Hoboken, NJ.

Vij, A. and Shankari, K. (2015) When is big data big enough? Implications of using GPS-based surveys for travel demand analysis. Transportation Research Part C: Emerging Technologies 56, 446–462.

Zhong, C., Arisona, S, Huang, X., Batty, M. and Schmitt, G. (2014) Detecting the dynamics of urban structure through spatial network analysis. *International Journal of Geographical Information Science*, 28, 11, 2178–2199.

Biographies

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.