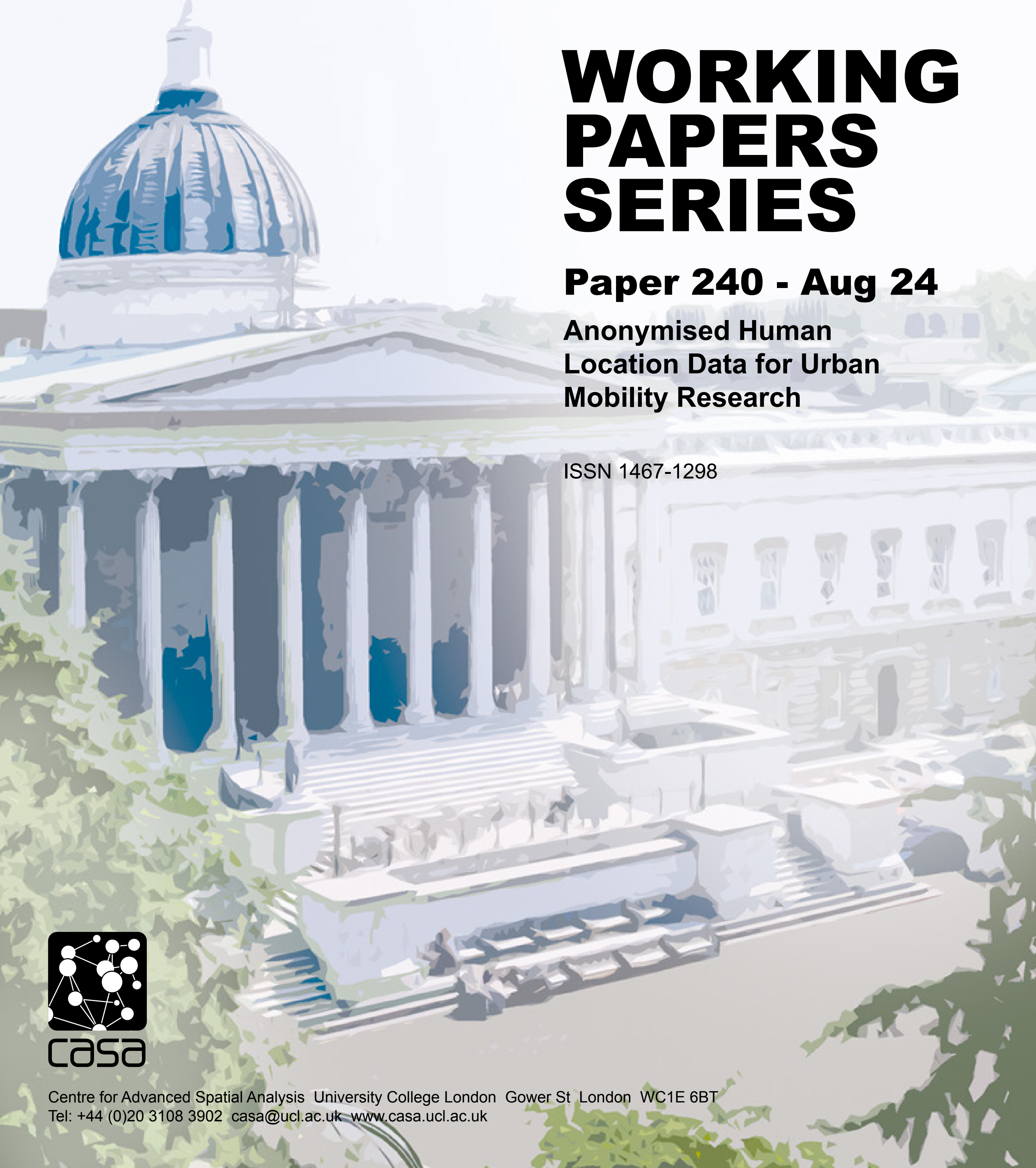
UCL CENTRE FOR ADVANCED SPATIAL ANALYSIS



WORKING PAPERS SERIES

Paper 240 - Aug 24

Anonymised Human Location Data for Urban Mobility Research

ISSN 1467-1298

casa

Centre for Advanced Spatial Analysis University College London Gower St London WC1E 6BT
Tel: +44 (0)20 3108 3902 casa@ucl.ac.uk www.casa.ucl.ac.uk

Anonymised human location data for urban mobility research

Chen Zhong*, Nilufer Sari Aslam, Yikang Wang, Zhengzi Zhou and Adham Enaya

Centre for Advanced Spatial Analysis, University College London,
90 Tottenham Court Road, London W1T 4TJ, UK
c.zhong@ucl.ac.uk

14 August 2024

Abstract.

Understanding human mobility is crucial for every aspect of daily life and the functioning of cities. Advanced by sensor technology and the big data economy, a highly influential body of research and applications on human mobility is driven by analyses of massive human location datasets, such as social media data and spending data. New data is emerging as rapidly as evolutionary technologies. Mobile app data is relatively new and has become available only in the recent decade. The derived data products are similar to those mainstreaming existing ones, mainly in trip-activity chains, counts, flow matrices, and derived indicators. However, the data bias varies across areas, periods and policy restrictions, requiring tailored data processing and validation solutions, which are not fully transparently discussed. This study contributes as a handbook for processing similar types of location points data, detailing engineering workflow and multi-stage validation techniques. Second, we present insights into the limitations and potential of data applications that tolerate the inevitable data bias. Finally, open trajectory and matrix data are shared for research purposes. The team will keep updating the methodology and results with the latest developments on GitHub.

Background & Summary

As demand for mobile services has multiplied over the last decade [1], the market is expected to continue growing rapidly with people wishing to access good quality mobile services wherever they live, work and travel. Simultaneously, mobile services generate an accurate and large amount of data from various sensors. Mobile app data is one of these emerging data by-products. Mobile app data are collected through software applications ('apps') that can be installed by the user on a smartphone and other wearable devices [2]. It comes with most minor ethical concerns as it is obtained with app users' consent. The most comparable data product on the market is mobile signal data (known as call detail records data), which mobile operators collect without users' consent. It has often been criticised for lacking a universal guidance and governance framework [3]. Due to concerns about privacy protection, access to mobile signal data and the derived products is limited. The commercial products are mostly aggregated, without transparent information on technical details. This greatly limits the employment of the data in education, research, and industrial uses. Mobile app data, as an alternative, overcomes most of the limitations. Mobile app data is nowhere near perfect, and there are different types of data uncertainty rooted in how it is generated. The bias may come from the overage of smartphone users, the market share of apps, and user preferences (e.g., the choice of the app, usage, and whether to give consent for location tracking). The data quality varies from one area to another. However, the bias issues, how we can handle it, and how much we can stand it have not been fully addressed.

Mobile in-app data has been explored in several studies to investigate mobility and activity patterns [4, 5], socio-spatial inequalities [6], disaster management [7], urban and regional development [8], economic activities [9], informing public health policy [10] and generic laws and mechanism of cities [11]. Apart from the research community, data has become a valuable asset in the digital economy, characterised by growing markets. Mainstream data companies, such as SafeGraph, CARTO, and Cuebiq, offer data products in a rather generic form and still focus on conventional points of interest (POI) data. Emerging location data products have been explored in recent years. However, due to the lack of transparent technical notes and a comprehensive understanding of data bias, immature business models and sometimes, regulation by GDPR, these types of data products are not easily recognised by the market.

Given the overall landscape of mobile in-app data research and commercial innovation, it is time to explore its full potential for human mobility application, with a clear awareness of its limits. The open science movement has significantly changed the research culture and facilitated an open and collaborative scientific community for information sharing and collective efforts, particularly during the pandemic. Open data sets are generated from mobile app data. For instance, daily time-series of three different aggregated mobility metrics in Italy were shared for monitoring the impact of the lockdown [12]; multiscale origin-to-destination (O-D) population flows across the US has been published for monitoring epidemic spreading [13]; a city-scale and longitudinal dataset of anonymised human mobility trajectories in Japan has been shared for benchmarking human mobility predicting models [14]. Over 11 billion geolocated cell phone records from Greater Mexico City were analysed dynamics before, during, and after COVID-19 [15]. In the UK, time-series counts data are made open at an aggregated level through national research facilities, e.g., the Consumer Data Research Centre (https://www.cdrc.ac.uk/about/). To promote the research community around mobile data for urban mobility studies, we introduce an openly available dataset at a finer level that provides anonymised trajectory data in the Greater London Area (GLA) and national O-D data at a fine geographical scale with the consideration of data bias and usability, and potential applications for urban mobility. To promote the research community around mobile data for urban mobility studies, we introduce an openly available dataset at a finer level that provides anonymised trajectory data in the Greater London Area and national O-D data at a fine geographical scale with consideration of data bias and usability, and potential applications for urban mobility.

Materials

The raw data used as input are from a UK-based data service company – Locomizer, who license mobile GPS data sourced from 200 mobile apps and pre-processed data (e.g., anonymisation using a cryptographic hash function, filtering noisy points and aggregation) to ensure the data adheres to local privacy regulations such as GDPR and contractual obligations with the suppliers. Although the information generated by user interactions with mobile applications could be rich, including user behaviour and device information, the data we used is simple and kept minimal attributes to minimise ethical risks and to be in line with other primary streaming automatic human mobility data (e.g., smart card data, tweets).

We used the data from November 2021 to demonstrate the workflow and share regenerated data sets for research use. We purposely selected the month after COVID-19 measures were completely lifted. It is not a holiday season and has no bank holidays. However, the Ultra Low Emission Zone (ULEZ) expansion was implemented on 25 Oct 2021, which may bring some unknown variability to the mobility patterns. Overall, the mobile app data in November covers about 1.028% of the population with variables across areas. i.e., 1.021% in the Great London Authority (GLA). The total number of recorded active devices (equivalent to users in this paper) with a significant number of points for activity identification and home and work detection is 793,502, and 628,649 (80%) users have records of at least 14 days. Some basic statistics about user counts, active user counts, and signal counts are summarised in Appendix C. The original location points data has minimal attributes, including anonymised device ID, latitude, longitude, and time tag.

Methods

Location data is a widely researched area with input from various domains (e.g., computer science, geography, urban planning, physics). The terminologies used may have different meanings in varied domain contexts. Therefore, we have defined the terminologies in Appendix B. Based on that, the overall data processing framework is presented in Figure 1, which illustrates detailed steps from original data sets to regenerated sharable data sets. The first step is to extract stay points by aggregating over-sampled noisy points. The second step combines spatial contextual information (i.e., POIs) and temporal patterns (e.g., visiting time) to infer activity types. It is possible to infer travel models by matching transition points with road networks. However, this is not scoped in this paper and is left for future updates. The final step summarises the processed data into commonly used data forms, e.g., OD flow matrix, trajectories and counts. Each intermediate result was validated with details presented in the later section of technical validation.

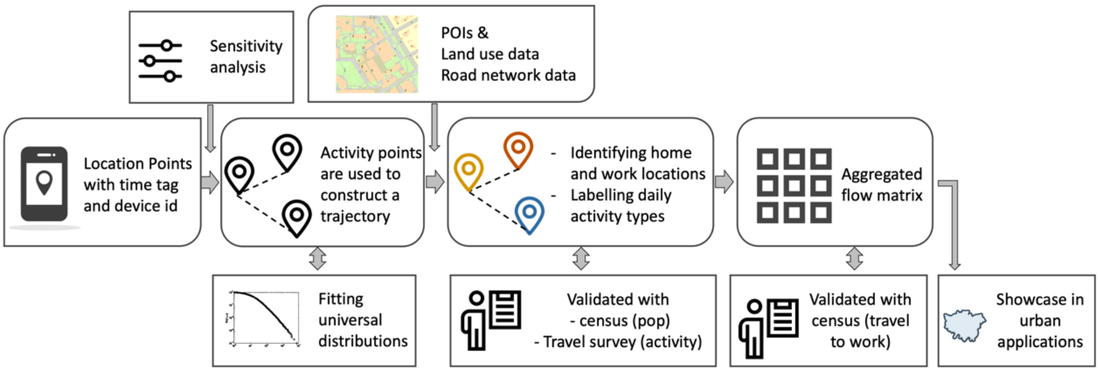

Sensitivity analysis
POIs &
Land use data
Road network data
Location Points with time tag and device id
Activity points are used to construct a trajectory
- Identifying home and work locations
- Labelling daily activity types
Aggregated flow matrix
Fitting universal distributions
Validated with
- census (pop)
- Travel survey (activity)
Validated with census (travel to work)
Showcase in urban applications

Figure 1. The overall data processing workflow.

Extracting stay points and activity locations by clustering methods and sensitivity analysis**:** Extracting stay points removes oversampled noisy points while maintaining minimal points for trajectory representation. The most used method is spatial clustering of sequential points. We eventually adopted the Infostop Python package [16], seeing its advantages in simplicity and computation efficiency in its fast C++ module. Infostop is a generic clustering methods-based framework to transform dense and rich location time series into sequences of events. In our context, events are equivalent to activities. While Infostop is effective in extracting stay points, the effectiveness is still largely affected by some self-defined critical parameters, i.e., stay as periods when an individual does not stray further than a **maximum distance D_max** for a **minimum duration t_min**. Previous works use spatial clustering or rule-based methods to set distance thresholds from 50 to 500m based on expert knowledge [17-21]. We propose incorporating a sensitivity analysis (shown in Appendix C) to determine the distance threshold rather than applying a universal setting.

Identify home and work locations: As implemented in most literature, we applied a simple rule-based classification to identify home and work locations. For each device ID, we take the stay point, recorded within a defined temporal window (i.e., 7 PM to 7 AM), which has the longest duration of stay compared to other location points and the most frequent visits as the home location. Similarly, a work location is identified from the non-home candidate locations for only weekdays with the longest stay (between 7 AM and 11 PM) and the most frequent visits.

Labelling activity types: We defined six routine activity types (i.e., home, work, education, eating and drinking, shopping type 1, shopping type 2, entertainment, and others) and labelled all detected stays to one activity each. Dividing shopping activities under two categories (i.e., frequently small consumptions and large infrequent consumptions) aligns with the literature due to behavioural differences [22]. Home and work activities are labelled for any stays at home and work locations by default. Other than home and work activities, they were labelled using joint probabilities of spatial and temporal features commonly used in the relevant literature [23, 24]. We factored spatial features by counting urban context around stay points, delineated by Points of interest (POI) collected from the Ordnance Survey (https://www.ordnancesurvey.co.uk/). In particular, a buffer area (d = 500 meters) is created around each stay point. The probability of attending certain categories of POIs (in Appendix D)is calculated by a Huff model [25]. Besides, **t**emporal features were counted as the probability of activity starting time. For each type of activity, we draw temporal signatures inspired by the literature [22] and customised by local travel surveys and time-use surveys (shown in Appendix E). The one with the highest joint probability is considered as the labelled activity.

Data Records and Usage

The dataset is available from the GitHub page (https://t.ly/dzlzB). This document could be considered technical notes and referenced when using the shared data. This dataset's records are composed of two main products: anonymised trajectories from 5000 randomly sampled users in the Greater London Area (GLA) and the national OD matrix at level 9 hexagon in the h3 geospatial indexing system (https://h3geo.org/) and the MSOA levels of the UK 2021 census geography (https://geoportal.statistics.gov.uk/).

Trajectory datasets contain the activity-trip-chain of 5,000 individuals who travelled across the GLA in November over 30 days. For ethical considerations. Each record (row) refers to an observation of a device (individual), which consists of the following columns:

- The device ID is the unique identifier of the mobile phone user
- Start time – is the timestamp of the observation sampled into 15-minute intervals.
- End time – is the timestamp of the observation sampled into 15-minute intervals
- Location - UK census tract – MSOA
- Activity label
- Duration – in minutes

O-D matrix data are provided in two files. One file contains travel-to-work trips only, and another includes all trips from all observations. Both O-D matrices are summarised at aggregated spatial units (i.e., hexagon and MSOA). The O-D matrix is in edge list format and contains three columns.

- Origin MSOA ID
- Destination MSOA ID
- Number of trips

Technical validation

Simple statistics of processed data

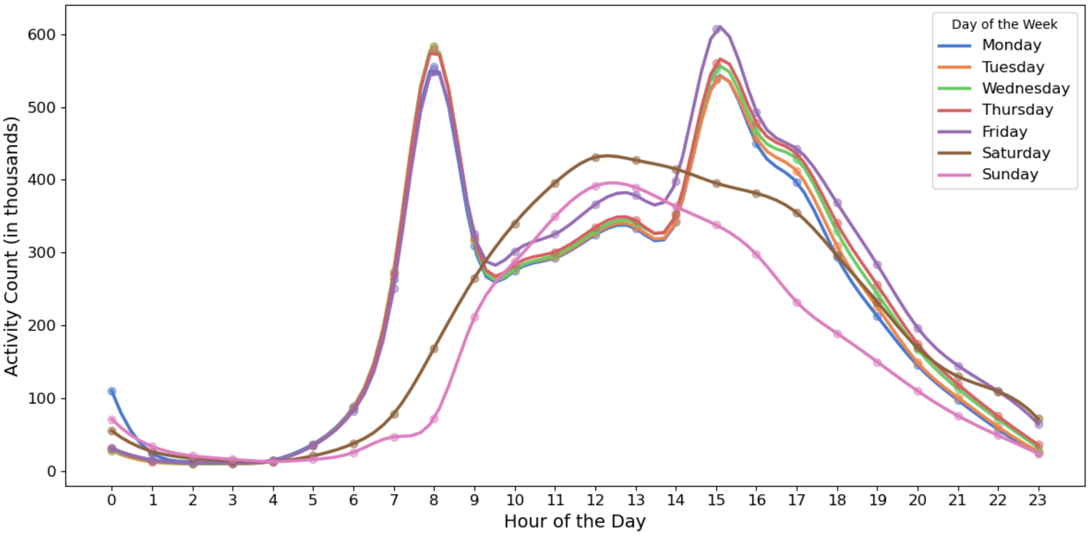


Figure 2. Average hourly activity counts in November 2021.

The above Figure 2 presents a simple statistical overview of processed data. The average counts of activities were summarised by starting time for each day of the week. All weekdays (Monday through Friday) show similar activity patterns with two peaks: one around 8-9 AM and another around 3-5 PM. In contrast, weekend activity patterns have one peak around midday. While Saturday presents slightly higher activity counts, the lowest activity counts are captured on Sunday counts throughout the day. The following three sub-sections present the validations introduced in Figure 1.

Distance distribution between each consecutive stays

Infostop generates stays by aggregating stationary points even though sensitivity analysis was performed to ensure that optimised parameters were set. We further validated that the generated stay points collectively follow the universal distribution. For this purpose, we measured the displacement distance between each consecutive stay, denoted as Δd. This distribution of displacement shall distinguish different types of diffusion processes, such as Lévy flights and random walk models. Although log-normal and exponential distributions have also been reported to fit well in specific datasets, the power-law distribution has proven more suitable for describing movement patterns over substantial distances. Pioneering studies utilising banknote tracking [26] and mobile phone call records (CDRs) [27] have demonstrated that a truncated power-law can approximate displacement distribution. In our analysis (plotted in Figure 3) the distances ranged from 1 km to 1000 km, and the fitted beta value of 1.24 aligns with the empirical range summarised in a detailed research review [11, 28]. This consistency underscores the robustness of our findings in characterising human movement patterns.

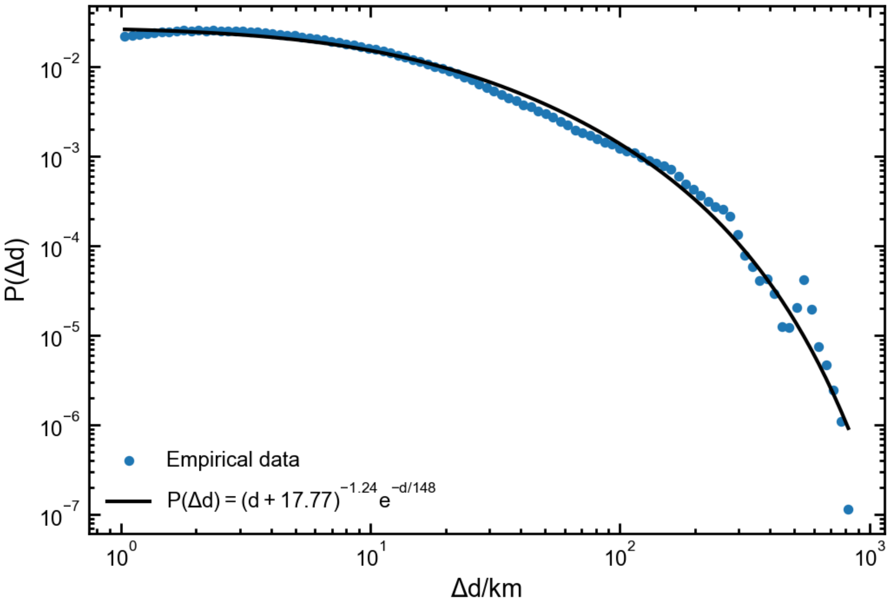


Figure 3. Displacement of activity locations fitting into a truncated power law

Correlation with usual residents, England: Census 2021

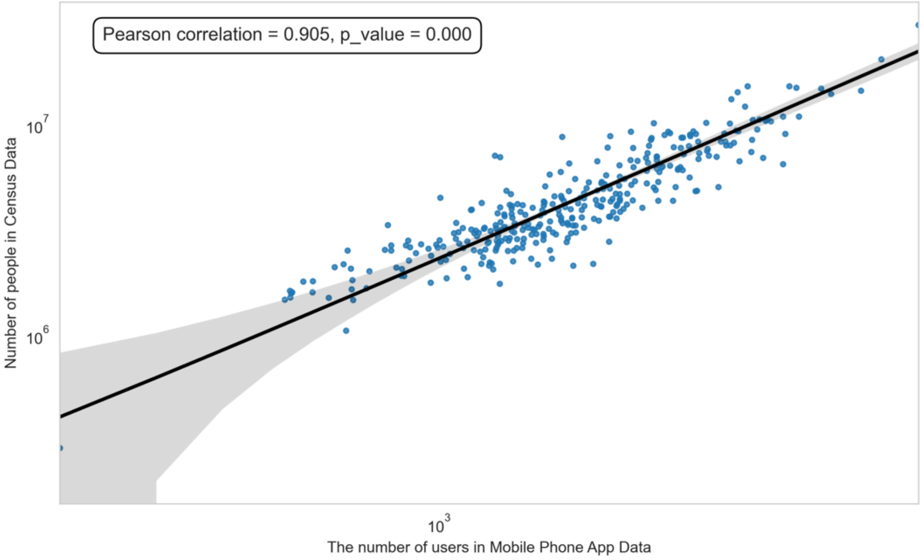


Figure 4. LAD-Level Correlation between the number of residents detected from Mobile Phone App Data and reported from Census Data

Figure 4 ***Figure 5*** illustrates the Local Authority Districts (LAD)-level Pearson correlation between two datasets, namely the number of users with home locations identified from the mobile app data and the number of usual residents estimated in the UK census 2021 (download from https://www.nomisweb.co.uk/sources/census_2021_bulk). The Pearson correlation coefficient is 0.905, indicating a decent representation. However, when we zoom into smaller areas, the correlation decreases

to 0.52 (shown in Appendix F). The limitation is somewhat anticipated. The detected home locations are based on simple rules, which makes it especially challenging to capture full working scenarios (e.g., night-time workers and mobile workers). Tuning the parameters used in the rule-based identification may slightly decrease/increase the numbers. Still, our sensitivity analysis shows no significant improvement. Appendix G has further discussed the issue and bias with statistics of hourly user counts of stay location type for each day of the week. The diversity and variability of working patterns have grown significantly in recent years, particularly post-COVID. Developing a comprehensive approach to identifying irregular home and working patterns will be one of the key topics in our improvements.

Correlation with travel to work, England: Census 2021

We compared our derived travel-to-work O-D matrix with multiple correlation measurements (Pearson, Spearman, and ratio as a measure of population penetration) at two levels – the Local Authority District (LAD) and the Middle layer Super Output Areas (MSOAs). The census travel-to-work data was collected through a combination of self-reported responses to specific questions related to commuting patterns and details about the time and distance. The Pearson correlation at the LAD level comparison is 0.95, which shows good data representativeness. For MSOA level validation, we took the entire England area but grouped MSOAs by upper-level LADs to understand the variabilities across areas. In total, 331 LADs were broken down into 7264 MSOAs. A log transformation is applied to avoid the impact of zero values and make the data near normal distribution for correlation analysis. As reported in the table, for the LDAs (311 areas out of a total of 331 LADs) with significant data records and more than one MSOA, the Pearson correlation shows decent results that range between 0.38 and 0.87 with an average of 0.7. Figure 5 provides further information about the spatial distribution.

Table 1. Statistics of correlation between LAD-level trips between MSOAs for both the census 2021 dataset and the mobile app dataset.

	Pearson	Spearman	Census (C)	Mobile (M)	Ratio (M/C)
mean	0.704	0.646	64944.511	1044.029	0.016
std	0.095	0.116	45228.264	760.316	0.005
50%	0.718	0.653	51323	790	0.016
75%	0.778	0.74	80649.5	1298.5	0.019
max	0.873	0.913	355566	5780	0.031

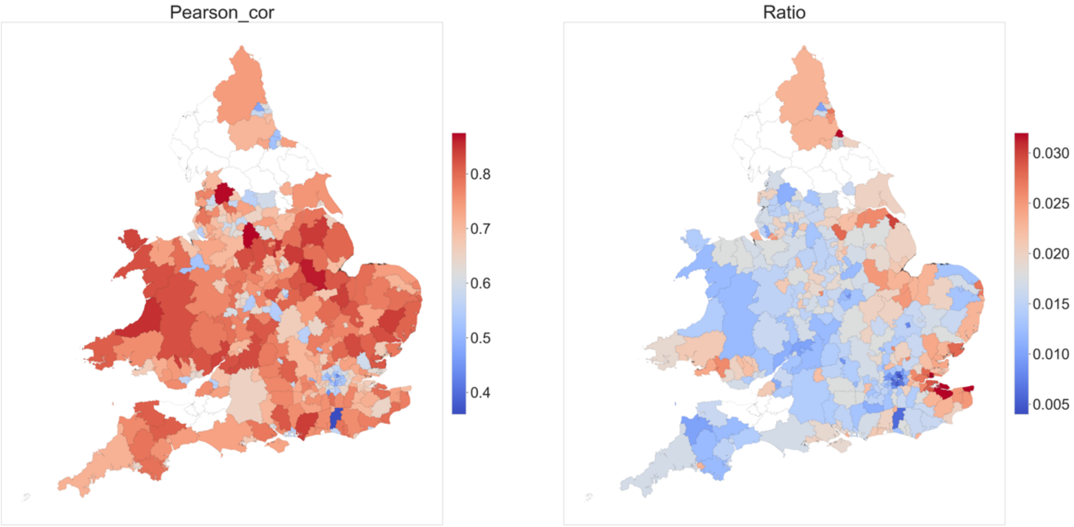


Figure 5. Pearson correlation and Ratio between travel-to-work trip detected from mobile app data and census data at LADs in England.

Validation in the context of example urban applications: spatial interaction model

The processed data, while not perfectly valid as documented, proves to be highly beneficial for various urban applications, especially at aggregated scales covering large areas. Here, we compared the travel-to-work data from the mobile app and census in the context of spatial interaction application. The four variants of the spatial interaction model were employed, including (unconstrained) gravity, production-constrained, attraction-constrained, and doubly constrained models. The model parameters (k: balancing factor, μ: production, α: attraction, β: for distance decay) were estimated, with two forms of distance decay functions (i.e., power and exponential) using two sets of travel-to-work data. We compared the performance of different models (reported in Figure 6) and found a very high correlation between the R^2 values estimated using both datasets. This suggests a strong relationship between the models' goodness of fit. The patterns observed in one dataset are mirrored in the other. This indicates that gravity models likely capture similar trip patterns across the datasets.

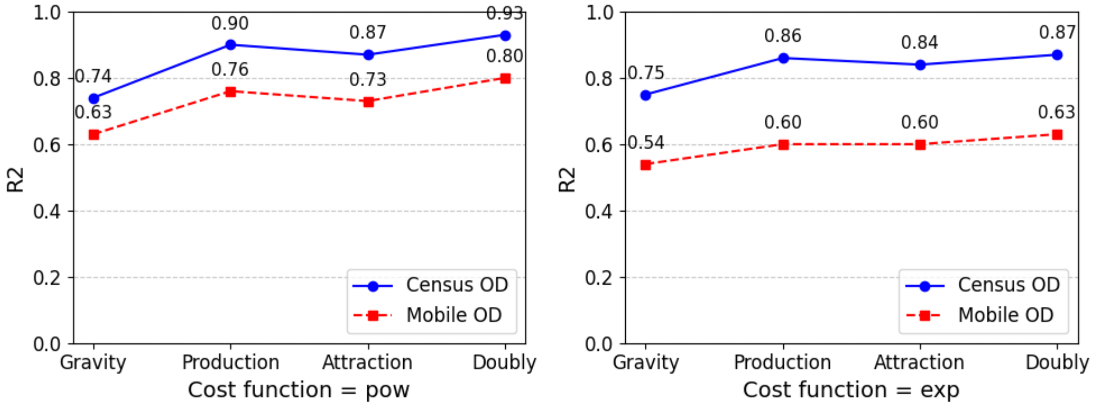


Figure 6. Plotting R^2 for different models presents the trends and consistency of goodness of fit across different models.

Validation in the context of example urban applications: spatial structure

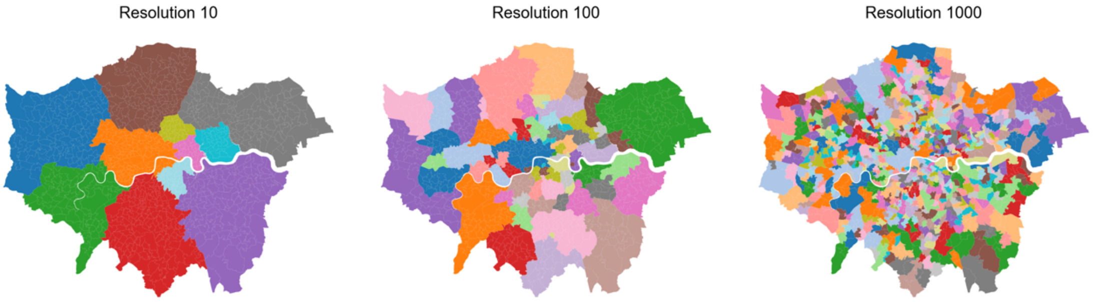


Figure 7 . Modularity-based community detection was applied to O-D data of all trips, delineating urban functional zones at different spatial scales.

Another commonly implemented application is to detect functional spatial structures based on flow data using community detection. In network science, a community refers to a sub-network that is dense internally and sparse externally [29]; revealing these communities allows us to understand the urban structure more intuitively. A further intuitive assumption is that urban networks are organised into hierarchically distinct communities, meaning that any given scale of community can be subdivided into smaller communities, which can be further subdivided, and so on [30]. One of the most widely used and arguably most universal methods is modularity maximisation. In recent years, modularity has been expanded to include a resolution parameter, which can be adjusted to discover communities at different scales [29]. Here, we demonstrate detected urban communities at three different resolutions (shown in Figure 7) The community detection results at different scales demonstrate the clustering of urban spatial units, which are largely explainable and overlapping with administration or social boundaries.

Conclusions and Future Directions

Automatically collected human location data has bias rooted in how it was generated, as mobile app data. When using the shared data sets, a user should bear in mind the limitations and also comply with the GDPR. First, a mobile app data point was collected whenever a user consented, location service was available, and an app was used. Different from these GPS trackers on vehicles for commercial purposes and continuously recording locations, the trajectories extracted from mobile app point data are meant to be incomplete. They should be considered sampled activities and trips from the sampled population. Mobility patterns or a complete travel diary should be extracted by further analysis. Second, our validation notes and demonstrated applications provide insights into the limitations and potential of data applications from the aspects of level of aggregation. More could be explored in the future. Third, to preserve privacy, we have aggregated the data to avoid any individuals' identification. We want to emphasise that sharing the data enables mobility analysis for planning purposes and can be used as benchmark data to compare with other cities, in a broader sense, to contribute to the culture of an open science community. We will continue this study dynamically and share our updates via our GitHub pages.

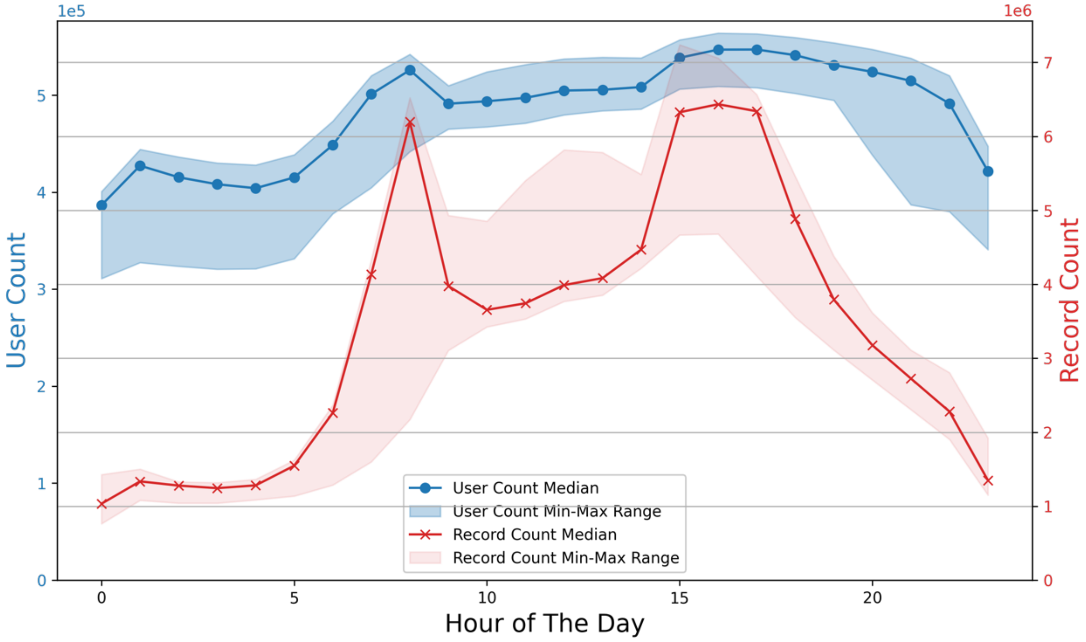
Ethical statement.

This project has received funding from the European Research Council (ERC) under the European

Union’s Horizon 2020 research and innovation programme (grant agreement No 949670). The study has been ethically approved by UCL's Research Ethics Committee (Ethics Application 21949/001)

Supplementary materials

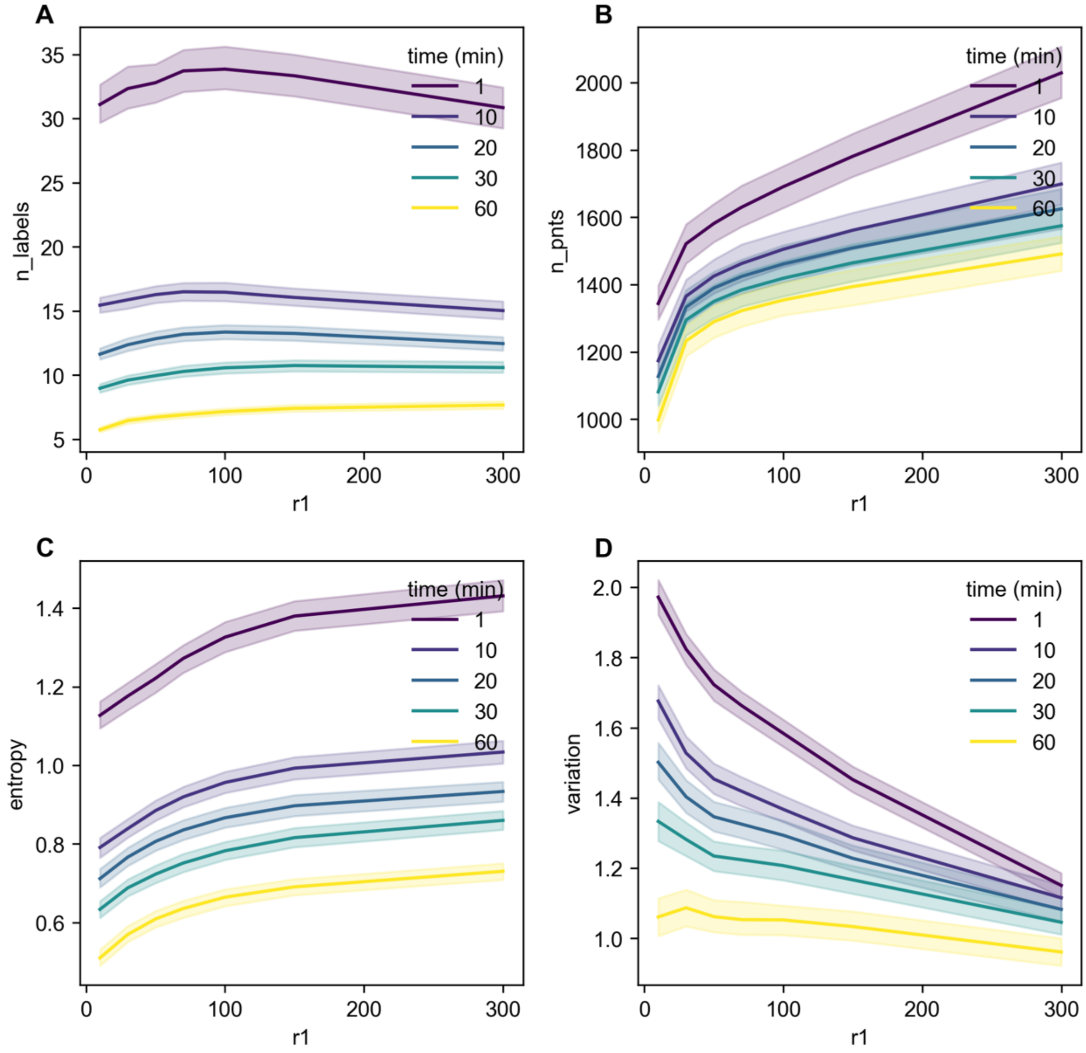
Appendix A Hourly number of device IDs (users) and points (records) in the dataset November 2021



Appendix B . Defining critical terminologies

- **Location points:** sometimes called events, are GPS points recorded as latitude and longitude in the raw mobile app data. Stay: is extracted from a series of stationary points; in other words, a stay is defined as a device remaining stationary for an extended period. In this work, a stay is equivalent to an activity (e.g., working for a few hours, staying at home overnight, having lunch in a restaurant, exercising in a gym).
- **Activity**: It is a stay labelled for travel purposes. Moving beyond the majority of the research focus on commuting patterns, we endeavoured to identify variable types of daily activities for a wider range of urban applications. Apart from primary activities, i.e., at home and work, we also labelled secondary activities, including education, eating and drinking, shopping (for regular daily grocery shopping, etc. and other shopping like outlets, etc.), entertainment, and others.
- **Trip**: is generated by connecting a series of consecutive non-stationary points. A trip means a move from one activity location to another associated with one or multiple travel purposes.
- **Trip-activity chain**: A series of short trips linked together between activity locations, such as a trip that leaves home, stops to drop off a kid at school, and continues to work. In the context of this work, a trip-activity chain is considered equivalent to a trajectory.
- **Origin-destination matrix (O-D matrix)**: a matrix summarising counts of trips between defined spatial units (e.g., census tracts).

Appendix C . Sensitivity analysis of parameters used for clustering



N_labels: The number of unique stops for each user

N_counts: The number of stops for each user

Entropy: The uncertainty of the number of stays for each user at different locations.

Variation: The variation of the number of stays for each user at different locations.

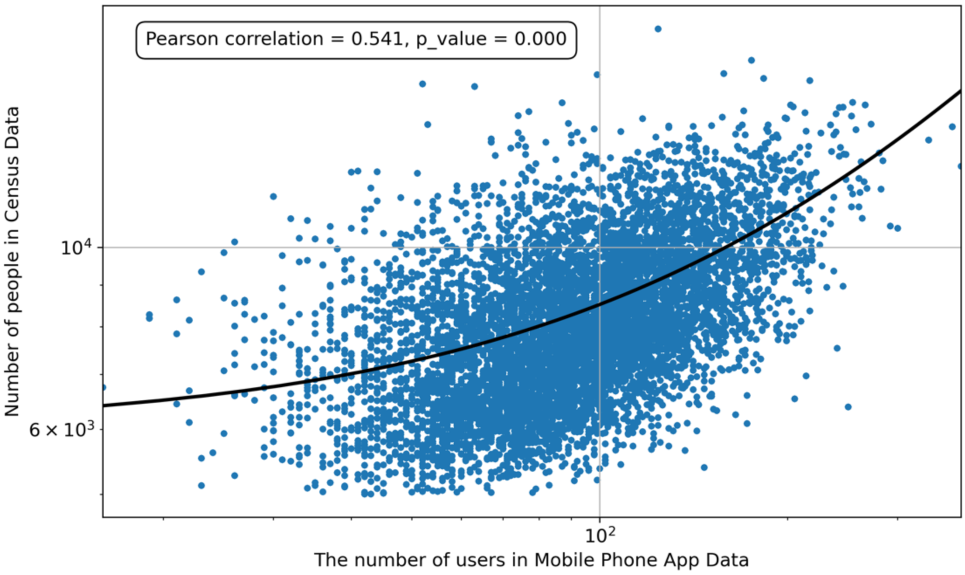
Appendix D. Categories of POIs used in activity labelling

Activity Types	Activity Location Type
Education	Primary, secondary, and infant schools, independent and preparatory schools, higher education establishments, other schools such as diving schools, drama schools, language schools, ballet and dance schools, beauty and hairdressing schools, etc.
Eating and Drinking	Restaurants, cafes, snack bars, tea rooms, pubs, bars, fish and chip shops, fast food delivery services, etc.
Shopping_type1	Grocers, markets, supermarket chains, Cash and carry, fishmongers, bakeries, etc.
Shopping_type2	Clothing, footwear, jewellery and fashion accessories, books and maps, florists, furniture, lighting, Electrical goods and components, second hand vehicles, etc.
Entertainment	Theatre, cinema, recreational, gambling, sport and entertainment services such as gym, etc.
Others	The rest of the POIs such as sport and entertainment, health, transport, etc.

Appendix E. Probability table used for proxy temporal signal of activities

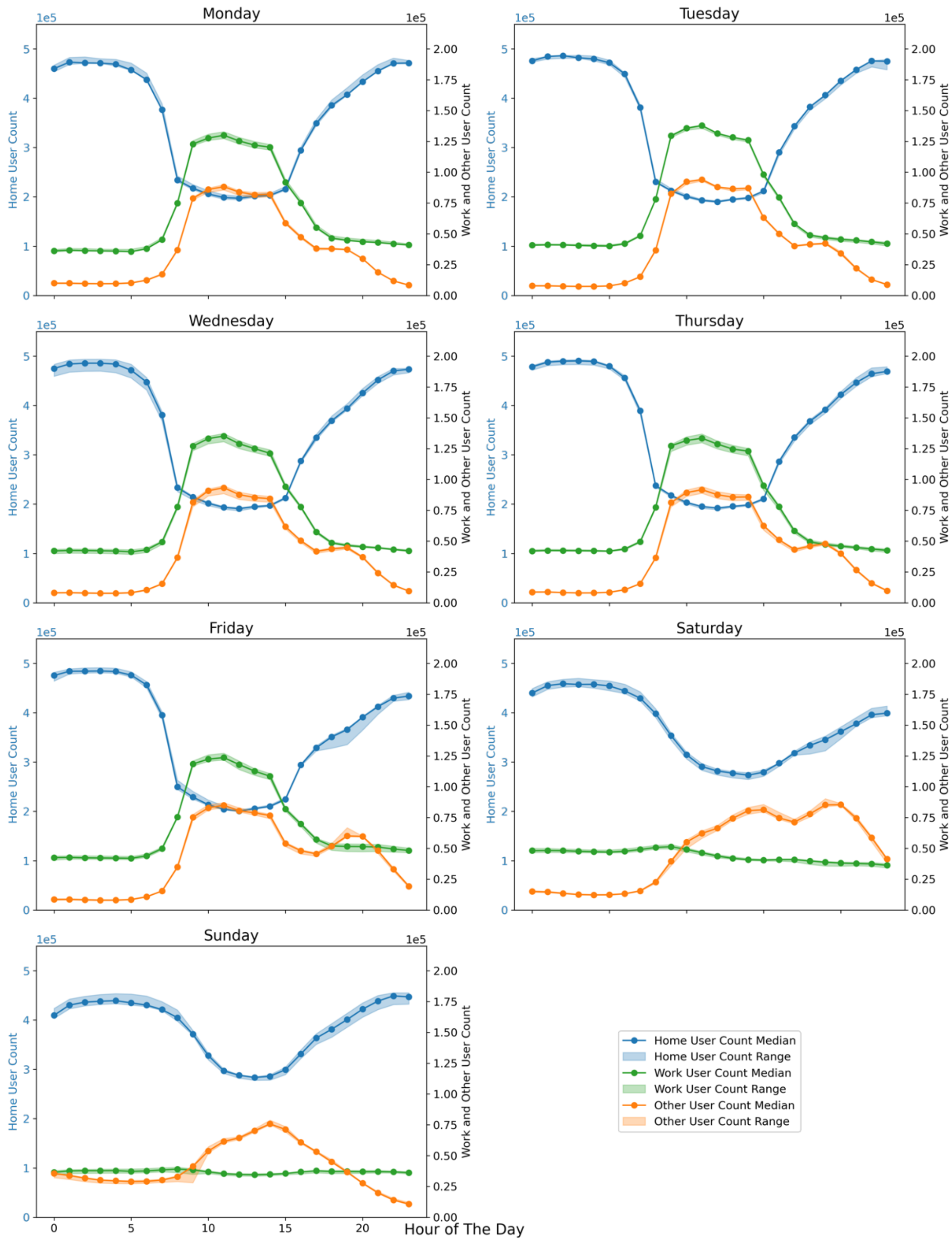
Start time		Education	Eating and Drinking	Shopping1	Shopping2	Entertainment	Others
06:00:00	07:00:00	0.1	0.1	0.1	0.1	0.1	0.3
07:00:00	08:00:00	0.7	0.7	0.5	0.1	0.1	0.3
08:00:00	09:00:00	0.9	0.7	0.5	0.1	0.1	0.7
09:00:00	10:00:00	0.5	0.5	0.5	0.5	0.3	0.5
10:00:00	11:00:00	0.3	0.5	0.5	0.7	0.5	0.7
11:00:00	12:00:00	0.3	0.9	0.5	0.7	0.5	0.7
12:00:00	13:00:00	0.3	0.9	0.7	0.3	0.7	0.3
13:00:00	14:00:00	0.3	0.7	0.7	0.5	0.5	0.3
14:00:00	15:00:00	0.3	0.5	0.5	0.7	0.5	0.5
15:00:00	16:00:00	0.5	0.3	0.5	0.7	0.5	0.5
16:00:00	17:00:00	0.3	0.5	0.5	0.7	0.5	0.5
17:00:00	18:00:00	0.3	0.7	0.5	0.7	0.5	0.5
18:00:00	19:00:00	0.3	0.9	0.3	0.1	0.3	0.3
19:00:00	20:00:00	0.3	0.9	0.7	0.1	0.7	0.3
20:00:00	21:00:00	0.1	0.5	0.7	0.1	0.9	0.5
21:00:00	22:00:00	0.1	0.5	0.5	0.1	0.7	0.5
22:00:00	23:00:00	0.1	0.3	0.5	0.1	0.5	0.5
23:00:00	24:00:00	0.1	0.3	0.5	0.1	0.3	0.5

Appendix F. MSOA-Level Correlation between the number of residents detected from Mobile Phone App Data and reported from Census Data**.**



Modifying the rules may increase the results with loose conditions/rules and decrease them with tight conditions/rules for the analysis. Within this mind, in our analysis, 109,817 records from the total of 757,811 individuals' records do not match with the census data due to 1) Unidentified home locations (7,677 individuals (1.01%)), 2) Unidentified work locations (32,618 individuals (4.3 %)) and 3) Unidentified home and work locations (42,279 individuals (5.5%)). Besides, 2.5% of individuals' anchor locations (27,243 records) do not match with MSOA spatial units due to various reasons such as border locations, etc.

Appendix G. Hourly user counts of stay location type for each day of the week.

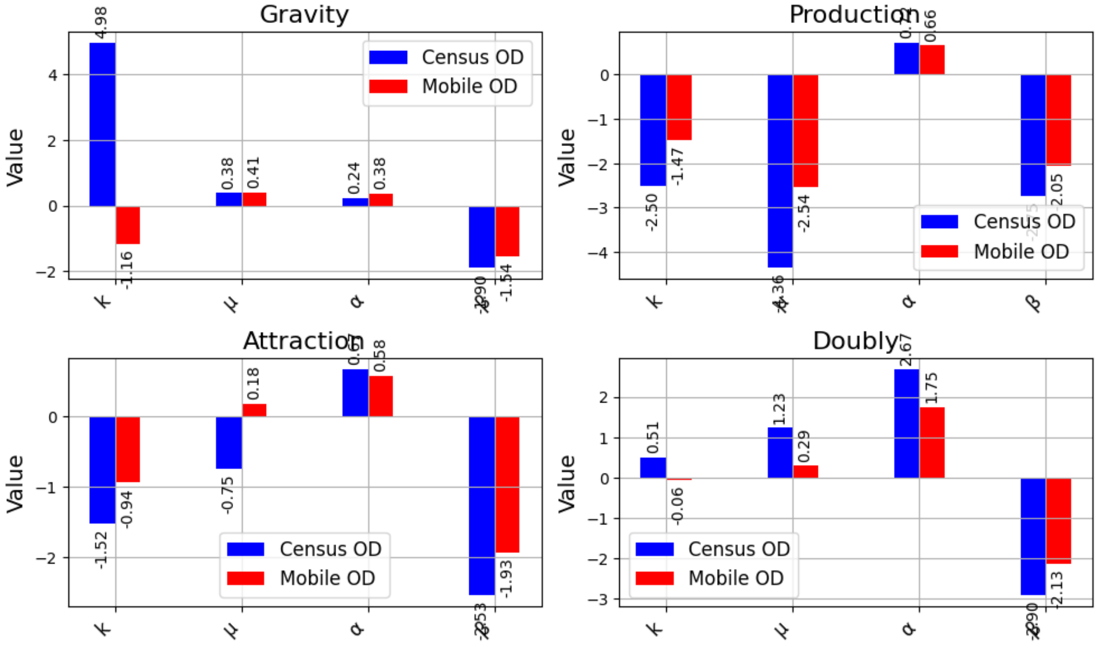


We calculated the user count by the type of stay locations. A user is considered to be staying at a location in a given hour only if this user has a stationary activity that covers a continuous period of at least 30 minutes in that hour. Therefore, a user can only be counted to a single type of location in each hour. Because the start and end time of activity requires a record, the first and last two days of the month (i.e., Monday 1st, Tuesday 2nd, Monday 29th, and Tuesday 30th of November) were removed from these statistics as they have lower user counts, especially for overnight stays at home location.

The plot illustrates the hourly distribution of user stay types across different days of the week, segmented into home, work, and other categories. Each subplot represents a specific day of the week, from Monday to Sunday, displaying the median values and the min-max range for the number of users in each category. The blue lines and shaded areas denote the median and range for home user counts, respectively. Similarly, the green and orange lines, along with their corresponding shaded areas, represent the median and range for work and other user counts.

From the plots, it is evident that home user counts exhibit a distinct diurnal pattern, peaking during the early morning and late evening hours while dipping during typical work hours. Conversely, work user counts show an inverse relationship, with higher counts during standard working hours between 9 AM and 3 PM on weekdays. We also observed significant night and weekend workers. The 'other' user counts demonstrate a more varied pattern, with noticeable peaks during midday on weekdays and early afternoon hours, Friday evenings, weekend afternoons, and Saturday evenings. Overnight stays at other locations are significant on Saturday nights.

Appendix H. Fitted parameters of spatial interaction models using travel-to-work OD matrix from census and mobile app**.**



Examining the estimated parameters for each model, there is generally consistency in the signs of the coefficients. Importantly, we observed a negative distance decay value (β), which aligns with the literature on gravity modelling. This likely suggests an inverse relationship between the number of trips between regions and the distance separating them. There are significant differences in some coefficients, such as k (balance factor) in the model. A statistical hypothesis could be conducted to determine if there is significant evidence to reject the null hypothesis (no difference between coefficients from models from both datasets). However, conducting such tests is beyond the scope of this paper, as it heavily depends on the application context. Careful selection of variables and datasets is crucial in structuring the model. In our case, we assumed job population in employment from census data is sufficient to model trips from mobile data. This strong assumption could explain some discrepancies in the coefficient magnitude observed.

References

1. Ofcom UK, *Meeting future demand for mobile data*. 2022.
2. Trasberg, T. and J. Cheshire, *Spatial and social disparities in the decline of activities during the COVID-19 lockdown in Greater London*. Urban Studies, 2021: p. 00420980211040409.
3. Jansen, R., et al., *Guiding principles to maintain public trust in the use of mobile operator data for policy purposes*. Data & Policy, 2021. **3**: p. e24.
4. Ross, S., et al., *Household visitation during the COVID-19 pandemic*. Scientific reports, 2021. **11**(1): p. 22871.
5. Trasberg, T. and J. Cheshire, *Spatial and social disparities in the decline of activities during the COVID-19 lockdown in Greater London*. Urban Studies, 2023. **60**(8): p. 1427-1447.
6. Gao, Q.-L., C. Zhong, and Y. Wang, *Unpacking urban scaling and socio-spatial inequalities in mobility: Evidence from England*. Environment and Planning B: Urban Analytics and City Science, 2024: p. 23998083241234137.
7. Yabe, T., et al., *Mobile phone location data for disasters: A review from natural hazards and epidemics*. Computers, Environment and Urban Systems, 2022. **94**: p. 101777.
8. Zhang, B., et al., *Delineating urban functional zones using mobile phone data: A case study of cross-boundary integration in Shenzhen-Dongguan-Huizhou area*. Computers, Environment and Urban Systems, 2022. **98**: p. 101872.
9. Dong, L., et al., *Measuring economic activity in China with mobile big data*. EPJ Data Science, 2017. **6**: p. 1-17.
10. Oliver, N., et al., *Mobile phone data for informing public health actions across the COVID-19 pandemic life cycle*. 2020, American Association for the Advancement of Science. p. eabc0764.
11. Alessandretti, L., U. Aslak, and S. Lehmann, *The scales of human mobility*. Nature, 2020. **587**(7834): p. 402-407.
12. Pepe, E., et al., *COVID-19 outbreak response, a dataset to assess mobility changes in Italy following national lockdown*. Scientific Data, 2020. **7**(1): p. 230.
13. Kang, Y., et al., *Multiscale dynamic human mobility flow dataset in the US during the COVID-19 epidemic*. Scientific data, 2020. **7**(1): p. 390.
14. Yabe, T., et al., *YJMob100K: City-scale and longitudinal dataset of anonymized human mobility trajectories*. Scientific Data, 2024. **11**(1): p. 397.
15. Flores-Garrido, M., et al., *Mobility networks in Greater Mexico City*. Scientific Data, 2024. **11**(1): p. 84.
16. Aslak, U. and L. Alessandretti, *Infostop: scalable stop-location detection in multi-user mobility data*. arXiv preprint arXiv:2003.14370, 2020.
17. Kalatian, A. and Y. Shafahi. *Travel mode detection exploiting cellular network data*. in *MATEC Web of Conferences*. 2016. EDP Sciences.
18. Usyukov, V., *Methodology for identifying activities from GPS data streams*. Procedia Computer Science, 2017. **109**: p. 10-17.
19. Wolf, J., et al., *Eighty weeks of global positioning system traces: approaches to enriching trip information*. Transportation Research Record, 2004. **1870**(1): p. 46-54.
20. Yang, Y., et al., *Detecting home and work locations from mobile phone cellular signaling data*. Mobile Information Systems, 2021. **2021**: p. 1-13.
21. Yazdizadeh, A., Z. Patterson, and B. Farooq, *An automated approach from GPS traces to*

complete trip information. International Journal of Transportation Science and Technology, 2019. **8**(1): p. 82-100.

22. Huang, L., Q. Li, and Y. Yue. *Activity identification from GPS trajectories using spatial temporal POIs' attractiveness*. in *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on location based social networks*. 2010.
23. Yin, L., N. Lin, and Z. Zhao, *Mining Daily Activity Chains from Large-Scale Mobile Phone Location Data*. Cities, 2021. **109**: p. 103013.
24. Sari Aslam, N., et al., *ActivityNET: Neural networks to predict public transport trip purposes from individual smart card data and POIs*. Geo-Spatial Information Science, 2021. **24**(4): p. 711-721.
25. Huff, D.L., *A probabilistic analysis of shopping center trade areas*. Land economics, 1963. **39**(1): p. 81-90.
26. Brockmann, D., L. Hufnagel, and T. Geisel, *The scaling laws of human travel*. Nature, 2006. **439**(7075): p. 462-465.
27. Gonzalez, M.C., C.A. Hidalgo, and A.-L. Barabasi, *Understanding individual human mobility patterns*. nature, 2008. **453**(7196): p. 779-782.
28. Alessandretti, L., et al., *Multi-scale spatio-temporal analysis of human mobility*. PloS one, 2017. **12**(2): p. e0171686.
29. Reichardt, J. and S. Bornholdt, *Statistical mechanics of community detection*. Physical Review E—Statistical, Nonlinear, and Soft Matter Physics, 2006. **74**(1): p. 016110.
30. Hilgetag, C.C. and M.-T. Hütt, *Hierarchical modular brain connectivity is a stretch for criticality*. Trends in cognitive sciences, 2014. **18**(3): p. 114-115.