Managing Variability: A sensitivity analysis approach applied to mobile app data

Sari Aslam N[*1], Zhong C[†1] and Wang Y[‡1]

[1]Centre for Advanced Spatial Analysis (CASA), University College London (UCL)
Gower St, London, UK

GISRUK 2024

Summary

Mobile app data provides accurate spatial and temporal details to present human trajectories. Recognising the challenges posed by the inherent variability and large volume of data, the research refines data processing techniques through the application of sensitivity analysis, explicitly using distance parameters in activity detection. The mobility indicator, activity disintegration (AD), is introduced to present the effectiveness of the model using sequential, spatial, and temporal search from individual mobile app trajectories. The results demonstrate how AD varies in different thresholds and cities in urban mobility analysis.

KEYWORDS: Stay locations, activity identification, sensitivity analysis, urban mobility, mobile app data.

1 Introduction

Mobile app data derived from Global Navigation Satellite System (GNSS) technology provides a comprehensive insight into understanding urban mobility patterns (Nguyen et al., 2020). However, due to inherent variability in recording frequency, this data is often extensive, making it impractical for direct use in urban studies and thus requires efficient data processing.

Activity location (stay location) is the first step in mobile app data processing. Previous works use spatial clustering (Kalatian and Shafahi, 2016) or other rule-based methods (Fang et al., 2018) to share a critical pre-set parameter: a distance threshold. This value varies from research to research and ranges from 50 to 500m (Usyukov, 2017; Wolf et al., 2004; Yang et al., 2021; Yazdizadeh et al., 2019), etc. This threshold is usually determined by the authors' experience and not through a quantitative process due to the lack of robust methods to find the optimal threshold. Building further assumptions with unchecked thresholds may increase biases in complex urban systems. Optimising these indicators improves the precision and reliability of the data analysis in urban mobility studies.

This study aims to refine data processing steps by examining a mobility indicator, AD (the number of activities based on the same locations), while checking the positioning of the activity sequentially, spatially, and temporally. Besides, the study applied various datasets using different thresholds in

[*] n.aslam.11@ucl.ac.uk
[†] c.zhong@ucl.ac.uk
[‡] yikang.wang.21@ucl.ac.uk

different cities to address the challenge of aligning the data's inherent variability and reliability in urban mobility research.

2 Data and Method

2.1 Mobile App Data

Mobile app data in the UK contains more than half a million individuals per day. Randomly selected 10,000 individuals' daily trajectories from the UK and different cities such as London, Leeds, Glasgow, etc., are used for the data analysis. Each data point has location attributes, i.e., latitude and longitude, date, time, and anonymous user identifier. Table 1 compares the proportion of the city populations in the UK and the proportion of the user counts in the mobile app data. It is important to acknowledge that mobile app data may exhibit biases due to inherent variability in over- or under-sampling frequencies, which requires comprehensive data processing steps.

Table 1 The city populations in the UK versus the user counts in mobile app data.

Selected cities	The city populations (%)	The user proportion (%)
London	11.13	11.91
Leeds	1.19	1.11
Manchester	0.58	0.64
Glasgow	0.87	0.59
Liverpool	1.27	0.53
Bristol	0.91	0.60
Newcastle	0.28	0.34
Nottingham	1.08	1.31

2.2 Data pre-processing

Figure 1 demonstrates the data pre-processing steps, which create spatial coverage areas to define activity space using distance measures by calculating the gyration between consecutive data points. After sorting the data points by timestamp for each individual, first, the distances from the first data point(D_i) to the second (D_{i+1}), the first to the third (D_{i+2}), and so forth are calculated based on Haversine distance measurement. Two data points are considered in the same activity space if the distance does not exceed a pre-defined threshold. Contrarywise, a new activity space begins with the following data point as its first data point. Second, data points do not always coincide with the centre of the coverage area. The centre of the points is identified using median of the all the points While merging the rest of the data points in step 2, activity duration is also calculated using the first and the last data points in the coverage area. The calculated activity duration is used for labelling stationary and non-stationary activities in step 3 that we set the minimum activity duration for stationary activities at 10 minutes. As the last step, the same locations are identified using a distance matrix between stationary activities. If the calculated distance value between stationary activities is less than the defined distance threshold, the location is labelled with the same unique location identifiers as the same location. Nevertheless, the chosen threshold warrants additional analysis to delineate activity spaces accurately.

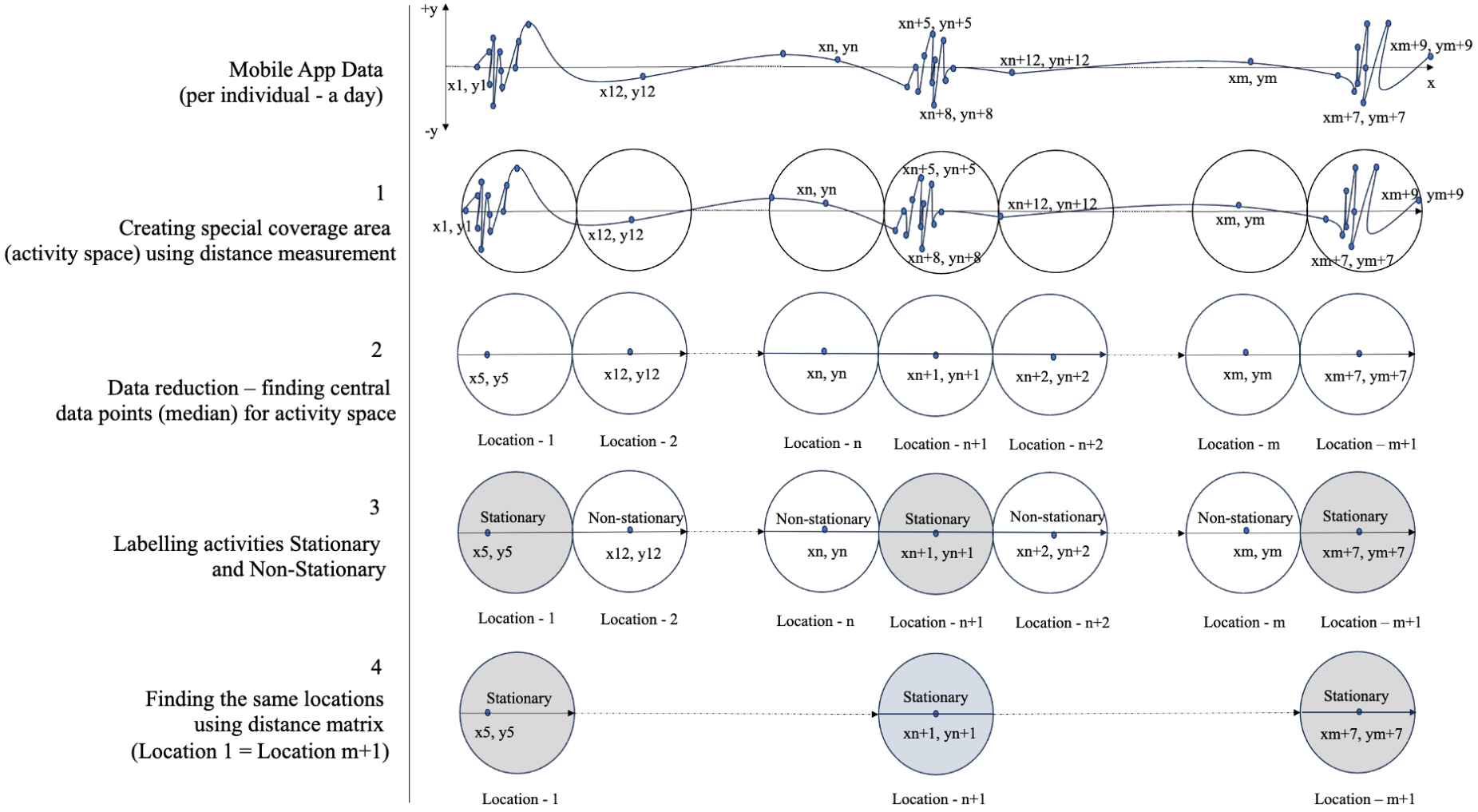

Figure 1 Schematic diagram of the activity generation process from mobile app data (drawn without scale).

2.3 Sensitivity Analysis

Sensitivity analysis determines different input variables that can influence a model's outcome. It is a way to assess the robustness of the results by systematically varying the threshold values to observe their impact on the data representation. In our application, a small distance threshold may divide the same activity into multiple activity spaces, referred to as activity disintegration. A large distance threshold may merge multiple activities into the same activity space. To prevent such scenarios, an optimum distance threshold is required that maximises activity identification while minimising such scenarios in the same locations.

2.4 Activity Disintegration

Activity disintegration (AD), as an urban mobility indicator, is defined as a count of activities at the same locations due to the inadequate threshold. The following criteria are applied and expected to be met to search ADs in the dataset: i) searching sequential proximity (the activities occur consecutively), ii) minimum duration (the combined duration of the two sequential activities should exceed 10 minutes, as 10 minutes is used as the minimum duration for an activity to be considered significant), and iii) spatial proximity (the activities occur in the same location). The same locations for stationary activities are identified in step 4). Besides, iv) the activity should not be a non-stationary activity characterised by a concise duration, typically 0 or 1 minute.

Multiple thresholds are applied to reduce AD, and the threshold that minimises the impact of AD needs to be selected for the robustness of the model. The following formula is used to calculate the percentage of AD in various countries in the UK:

$$AD\ (\%) = \left[\frac{AD_{count}}{A_{total}}\right]_{tk} \tag{1}$$

Where AD (%) refers to the percentage of counts, t and k refer to the number of thresholds and counties in the UK, respectively. AD (%) is calculated by dividing the AD counts by the total activities in each threshold based on counties in the UK.

3 Results

The set of criteria is applied to the 10000 users randomly picked from the UK mobile app dataset to find the optimal threshold for the activity identification process. Figure 2 illustrates the results of AD and the total number of activities based on the various thresholds. As we increase the distance threshold, AD, mainly due to a smaller threshold, starts to reduce and stabilise at some point (6% at 50 metres), similar to the total activity counts, i.e., 50 metres. The optimum spot for the threshold values, where the AD is captured at a minimum, is 6% of the 50-metre distance in Figure 2.

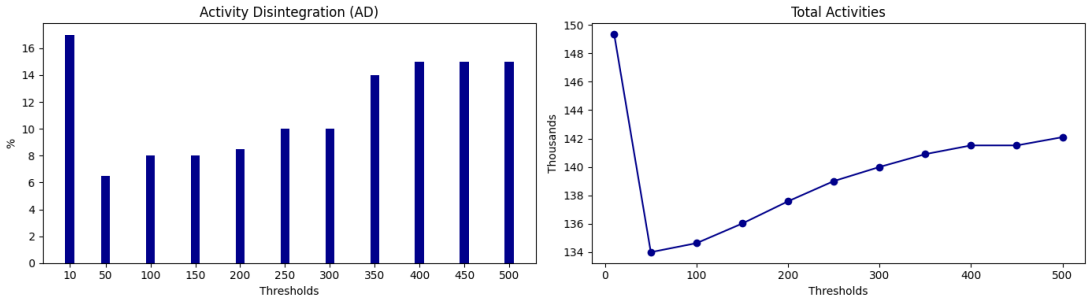

Figure 2 The proportion of AD and the total activity counts at different thresholds are derived from randomly selected individuals from the UK dataset.

Figure 3 illustrates AD counts at various thresholds for different cities in the UK, such as Greater London, Leeds, Manchester, Glasgow, Liverpool, Bristol, Newcastle, and Nottingham. The percentage of AD counts is calculated using randomly selected 10000 individuals for each city at thresholds ranging from 10 to 450. It can be seen from Figure 3 that each city has a unique pattern of AD counts across the thresholds, indicating differing levels of activity disintegration. While the minimum AD counts occur around the range of 50 to 150 m thresholds for London, Leeds and Manchester, Manchester shows the lowest AD counts, particularly at lower thresholds across the other cities. In addition, small cities like Bristol, Newcastle and Nottingham show less variability across thresholds in ADs as compared to medium-sized cities such as Liverpool, Glasgow, and Leeds.

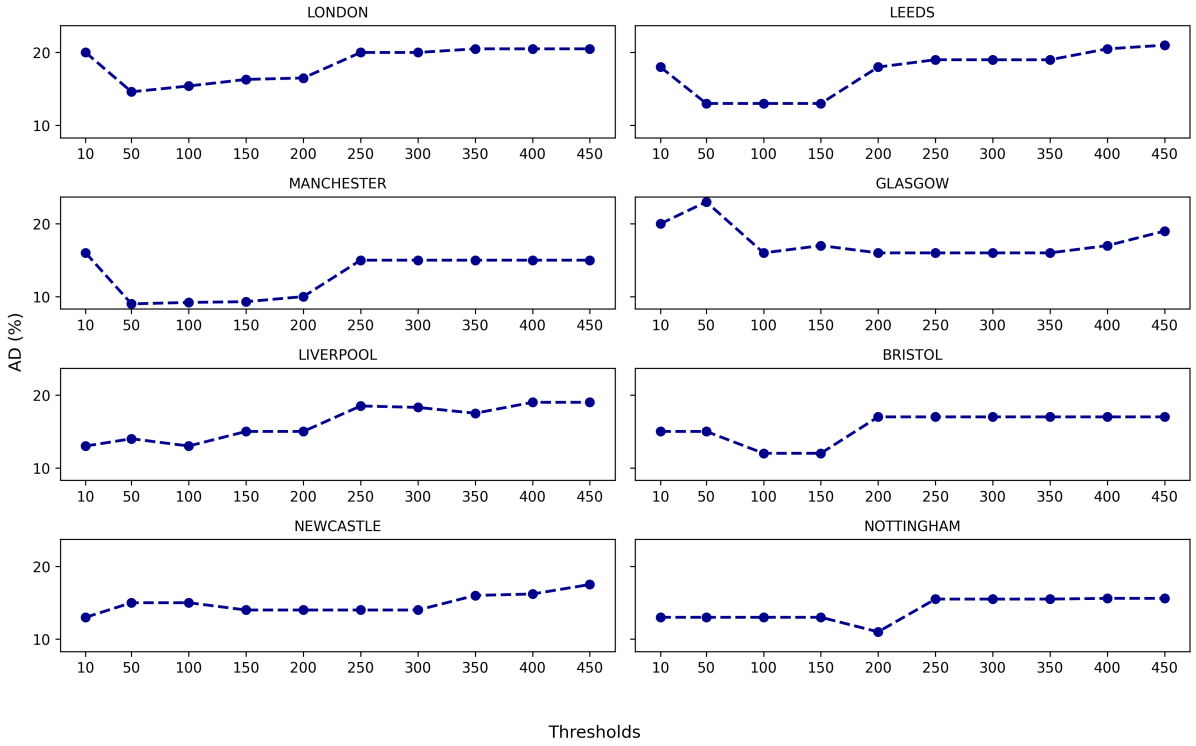

Figure 3 Activity Disintegration (AD) counts at various thresholds for different UK cities.

Moreover, some similarities in terms of AD are captured between Figure 2 and Greater London in Figure 3. For instance, the proportion of AD at a threshold of 10, initially over 16%, distinctly drops to approximately 6 to 7% at a threshold of 50. This level of AD maintains a steady state for nearly an additional 150 metres in both illustrations. This trend can be attributed to the fact that the population of Greater London accounts for about 13% of the UK dataset. Consequently, when users are selected randomly from this dataset, there is a higher likelihood of representing Greater London users than those from other cities.

4 Conclusion

This study demonstrates the utility of sensitivity analysis in threshold optimisation for mobile app data processing. It provides valuable insight into understanding the urban mobility indicator in various cities, particularly in how different thresholds impact the disintegration of activities. The results show that each city has a unique pattern of AD counts across the thresholds, indicating differing levels of activity disintegration. Despite certain limitations, such as specific datasets, the findings have important implications in optimising precise mobility data analysis workflow in urban studies. Future research will expand using sensitivity analysis with multi-variables in diverse urban environments and datasets.

References

Fang, Z., Jian-yu, L., Jin-jun, T., Xiao, W., Fei, G., 2018. Identifying activities and trips with GPS data. IET Intelligent Transport Systems 12, 884–890. https://doi.org/10.1049/iet-its.2017.0405.

Kalatian, A., Shafahi, Y., 2016. Travel Mode Detection Exploiting Cellular Network Data. MATEC Web Conf. 81, 03008. https://doi.org/10.1051/matecconf/20168103008.

Nguyen, M.H., Armoogum, J., Madre, J.-L., Garcia, C., 2020. Reviewing trip purpose imputation in GPS-based travel surveys. Journal of Traffic and Transportation Engineering (English Edition) 7, 395–412. https://doi.org/10.1016/j.jtte.2020.05.004.

Usyukov, V., 2017. Methodology for identifying activities from GPS data streams. Procedia Computer Science 109, 10–17. https://doi.org/10.1016/j.procs.2017.05.289.

Wolf, J., Schönfelder, S., Samaga, U., Oliveira, M., Axhausen, K.W., 2004. Eighty Weeks of Global Positioning System Traces: Approaches to Enriching Trip Information. Transportation Research Record 1870, 46–54. https://doi.org/10.3141/1870-06.

Yang, Y., Xiong, C., Zhuo, J., Cai, M., 2021. Detecting Home and Work Locations from Mobile Phone Cellular Signaling Data. Mobile Information Systems 2021, 1–13. https://doi.org/10.1155/2021/5546329.

Yazdizadeh, A., Patterson, Z., Farooq, B., 2019. An automated approach from GPS traces to complete trip information. International Journal of Transportation Science and Technology 8, 82–100. https://doi.org/10.1016/j.ijtst.2018.08.003

Acknowledgements

This work was supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (No. 949670), and it has been approved by the University College London Research Ethics Committee (project id: 21949/001).

Biographies

Dr Nilufer Sari Aslam is a research fellow in urban mobility and inequality at CASA, UCL. Her research interests are spatial-temporal data mining machine learning (ML) in urban mobility and transport planning.

Dr Chen Zhong is an associate professor of urban analytics at CASA, UCL. Her research interests lie in spatial data analysis, machine learning (ML), urban modelling, and data-driven urban and transport planning methods.

Yikang Wang is a PhD student at CASA, UCL. His research interests include human mobility analysis and spatial causal inference.