

## A high-precision heuristic model to detect home and work locations from smart card data

Nilufer Sari Aslam, Tao Cheng & James Cheshire

To cite this article: Nilufer Sari Aslam, Tao Cheng & James Cheshire (2018): A high-precision heuristic model to detect home and work locations from smart card data, Geo-spatial Information Science, DOI: [10.1080/10095020.2018.1545884](https://doi.org/10.1080/10095020.2018.1545884)

To link to this article: <https://doi.org/10.1080/10095020.2018.1545884>



© 2018 Wuhan University. Published by Informa UK Limited, trading as Taylor & Francis Group.

---



Published online: 28 Nov 2018.

---



Submit your article to this journal [↗](#)


---



View Crossmark data [↗](#)

---

# A high-precision heuristic model to detect home and work locations from smart card data

Nilufer Sari Aslam <sup>a</sup>, Tao Cheng <sup>a</sup> and James Cheshire <sup>b</sup>

<sup>a</sup>SpaceTimeLab, Department of Civil, Environmental and Geomatic Engineering, University College London, London, UK; <sup>b</sup>Department of Geography, University College London, London, UK

## ABSTRACT

Smart card-automated fare collection systems now routinely record large volumes of data comprising the origins and destinations of travelers. Processing and analyzing these data open new opportunities in urban modeling and travel behavior research. This study seeks to develop an accurate framework for the study of urban mobility from smart card data by developing a heuristic primary location model to identify the home and work locations. The model uses journey counts as an indicator of usage regularity, *visit-frequency* to identify activity locations for regular commuters, and *stay-time* for the classification of work and home locations and activities. London is taken as a case study, and the model results were validated against survey data from the London Travel Demand Survey and volunteer survey. Results demonstrate that the proposed model is able to detect meaningful home and work places with high precision. This study offers a new and cost-effective approach to travel behavior and demand research.

## ARTICLE HISTORY

Received 6 October 2017  
Accepted 8 October 2018

## KEYWORDS

Smart card data; activity location modeling; heuristic primary location model; home and work locations; human mobility pattern; urban activity pattern

## 1. Introduction

Traditionally, activity-based travel demand models rely on travel demand surveys. These surveys are designed to gather rich information about travel choices but are expensive to administer, and typically cover only a short time span and a relatively small proportion of travelers.

Recent advancements in information and communication technologies have renewed interest in the activity-based approaches to human behavior research and urban planning. Large datasets of the kind generated by smart card systems offer enormous potential to better represent human travel behavior (Bagchi and White 2005; Pelletier, Trépanier, and Morency 2011). Despite such data sources lacking specific demographic information or information regarding users' journey purposes, many of these aspects of human mobility can be inferred from smart card data when travel patterns are regular (Maat, van Wee, and Stead 2005; Manley, Zhong, and Batty 2018).

With this in mind, we focus here on the detection of primary locations and activities based on public transport smart card records for London (UK). The model uses journey counts as an indicator of usage regularity, *visit-frequency* to identify activity locations for commuters, and *stay-time* for the classification of work and home locations and activities. The results from the model were validated using responses from the London Travel Demand Survey (LTDS) and volunteer survey data.

We used a heuristic approach as the basis from which to model individual mobility patterns for a large urban center, which deviates from related studies that seek to estimate primary locations. The heuristic primary location model is based on individual smart card dataset available from London and involves:

- An approach to transport planning that can utilize large datasets of individual travel data thereby eliminating the need for expensive and time-consuming travel demand surveys.
- Combining the spatial and temporal attributes such as start station, end station, *visit-frequency*, and *stay-time* in a comprehensive model of human mobility for the identification of key user locations and activities with high precision.

This paper will first present a brief overview of related work conducted in the field before defining the methodology for its study. It will then present the data and describe the application of the method, using London as a case study. The final section presents conclusions and future directions this research might take.

## 2. Related work

Identifying individuals' primary locations is essential for the analysis of human mobility. When using

smart card data, this process typically falls into two categories: estimation of origin–destination (OD) matrices, and identification of primary locations such as home and work (Anda, Erath, and Fourie 2017; Zou et al. 2016).

The identification of the origin and destination of a journey is fundamental to any travel analysis (Alsger et al. 2016), but this information is not always complete; in many automated fare collection (AFC) systems, users “tap in” to begin their journeys without “tapping out” to complete it (Chakirov and Erath 2012). For example, in London’s Underground (London’s subway system), barriers require users to tap in and out, but bus travel requires only a single tap as users board. To account for this, Barry et al. (2002) proposed two pragmatic assumptions. The first assumption states that the majority of commuters return to the destination station of their first journey to commence their next journey. The second one states that a substantial number of people finish the last journey of the day at the station where they started their first journey of the day. Built on this hypothesis, many researchers have been able to infer OD estimates from smart card data. For example, Zhao, Rahbee, and Wilson (2007) calculated the OD matrices from an origin-only AFC system in Chicago, while numerous other studies have also focused on smart card data for OD estimation (Trépanier, Tranchant, and Chapleau 2007; Seaborn, Attanucci, and Wilson 2009).

Fewer studies have focused on the identification of primary locations using smart card data (Zou et al. 2016). Heuristic (rule-based) or statistical approaches have been proposed in the related literature (Luong 2015; Zou et al. 2016; Anda, Erath, and Fourie 2017). Of note to heuristic study is the study of Devillaine, Munizaga, and Trépanier (2012) in which researchers defined work activities in Santiago as those which lasted longer than 2 h (weekday) and longer than 5 h in Gatineau (weekday) for adult-registered cards. In the same paper, home locations were identified as the destination of the last journey of the day. On the other hand, Hasan et al. (2012) proposed a simple mobility model for predicting primary locations using the frequency that individual users visit locations in a city. The most frequently visited places were classified as the home locations, whereas the second-most-visited locations concentrated around city centers and were identified as the work locations. A more recent study by Zou et al. (2016) extended Barry et al.’s (2002) assumptions by using the travel distance of home-based journeys to identify users’ home station in their Beijing study area. Their center-point-based detection algorithm was only able to identify 88.7% of passengers’ home locations by mining one week of the card transaction data. The rule-based algorithm proposed in this paper makes

use of similar assumptions with the addition of visit-frequency to identify home and work stations with high precision in the London area by mining two months’ worth of smart card data.

A number of studies have proposed probabilistic models as an alternative to the rule-based approach to activity identification. One such study was carried out by Chakirov and Erath (2012), who examined detection of home- and work-related activities based on smart card data for public transport in Singapore. Although their study used a rule-based approach to identify work locations based on duration, it mainly focused on the discrete choice model with activity duration, start-time, and land use. The results were calibrated using information from a local travel diary survey, whilst information for land use was taken from the city’s official planning – Master Plan. Another probabilistic model approach was proposed by Li et al. (2015) who identified the most likely home and work pairs from smart card data based on duration and frequency. Using a rank aggregation technique and spectral analysis, the authors derived a comprehensive location ranking list in addition to identifying periodic travel patterns. The model was applied in Singapore, and the home location results were validated against the city’s urban planning dataset. Finally, Han and Sohn (2016) presented a probabilistic model derived from an unsupervised learning approach to identify activity patterns from smart card transactions. They proposed a continuous Hidden Markov Model that uses emissions probability to find eight clusters interpreted as patterns for home and out-of-home activities. An advantage of this model is that it can find the cluster membership of new observations, as well as generate activity chains to build simulation data. Although their model presents a way to discover activity patterns, the authors did not provide an indication of accuracy. Thus, the processing cost of such large amounts of smart card data is uncertain in this approach (Anda, Erath, and Fourie 2017; Zou et al. 2016).

Besides smart card data, there are studies predicting home and work locations from other data sources such as mobile phone call detail records, social media data, and GPS data (Ahas et al. 2010; Deng and Ji 2010; Isaacman et al. 2011; Hasan, Zhan, and Ukkusuri 2013; Jiang et al. 2016; Lotero et al. 2016). Due to privacy concerns, mobile phone and GPS data are not easy to obtain. Furthermore, social media data lacks the geotagged information for widespread and reliable identification of home and work locations.

In summary, our model uses a heuristic approach and it is expected that in the coming years with the rapid pace of technological advancements, data-centric rule-based and learning-based (machine learning) approaches will gain traction (Agard, Morency, and Trépanier 2007; Chakirov and Erath 2012; Zou et al.

2016; Alex, Saraswathy, and Isaac 2016). These approaches are better suited to leverage the wealth of data expected to be available. For this reason, a clear and simple framework is necessary to consistently identify human mobility patterns and improve transport planning (Hasan et al. 2012; Luong 2015).

### 3. Methodology

In our study, we have adopted four stages to detect the home and work locations from smart card data. Figure 1 illustrates these stages, described in further detail.

#### 3.1. Data preprocessing

Data preprocessing aims to improve the accuracy of the smart card data by cleansing the dataset. Hence, the useful records should be selected from the massive dataset by the application of the following rules:

Single journeys on any given day by individual users were excluded from the dataset, as they do not provide sufficient information about location (Chakirov and Erath 2012). Additionally, journeys where the origin and destination stations were the same do not provide any meaningful information for the analysis and therefore were excluded. Lastly, journeys without alighting time and location due to the missing tap-out information were also excluded from the dataset.

#### 3.2. Identify regular commuters

Regular commuters use the transport network for their daily home- and work-related travel. To establish the regularity of usage, journey count has been defined by the number of journeys carried out by each user. This separates frequent and regular users from sporadic users of the network. Too small a count will include a large number of irregular users

in the dataset, whilst too high a threshold will be too restrictive and leave out regular users from the analysis (Hasan et al. 2012). The appropriate journey count threshold will create a meaningful dataset for the study.

#### 3.3. Heuristic primary location model

The spatial and temporal regularity of journeys is a key driver in identifying mobility patterns. With the smart card data available, we have been able to use the OD of journeys to provide a spatial aspect, whilst the start-time, end-time, and *stay-time* of activities have given a temporal dimension for journeys.

In this paper, a *journey* is defined as one-way travel from one station to another. Additionally, *consecutive journeys* are defined as one journey after the other without any interruption such that the last stop of the first journey is the first stop of the second journey. The commuters carry out primary activities at the primary locations between two consecutive journeys.

The primary location model applied in this study used journey count and considered *visit-frequency* and *stay-time* as indicators in the model to identify home and work locations. *Visit-frequency* was defined as the number of times a specific user visited a location, whilst *stay-time* was defined as the activity duration between two consecutive journeys.

##### 3.3.1. Home location identifier

The algorithm selects the origin station of the first journey and the destination station of the last journey of the day for each user. If start and end stations match or are in close proximity (walking distance,  $\leq 500$  m), selected stations are analyzed further. The next stage is to pass these selected stations through the criteria of the *visit-frequency* threshold. If a station is identified more than the defined threshold for *visit-frequency*, it is classified as a home location for that user. It is

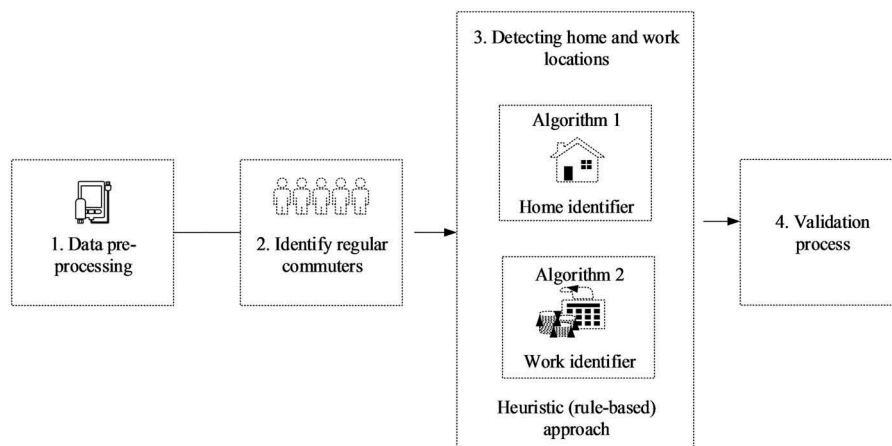


Figure 1. Workflow diagram of framework.

possible that the algorithm fails to highlight any station as a home location if none meet the expected criteria (Figure 2). At the same time, it is also possible that the algorithm may find more than one station that fits the criteria for a user's home location. In that case, Long, Zhang, and Cui (2012) and Wang et al. (2017) assigned the home station based on residential areas by using the land-use information. However, this does not apply well to cities with extensive transportation networks, as multiple stations identified may fall within residential areas.

Large cities are often not clearly segregated by land use; residential areas frequently contain work locations and other land-use types. Therefore, an approach based on distance and rank using frequency attributes provides a more meaningful outcome than dependence on land-use data.

This paper makes use of the assumption first made by Barry et al. (2002) that a high percentage of commuters end the last journey of their day at the same station where they started the first journey of their day. This location is significant as it represents the home location of the user. Although Zou et al. (2016) made a similar assumption, this paper combined the indicator of frequency threshold (Hasan et al. 2012) with the addition of distance indicator (walking distance,  $\leq 500$  m), to increase the accuracy of the findings.

### 3.3.2. Work location identifier

To identify the work locations, all consecutive journey pairs (J1 and J2) for all working days are evaluated. In the model, the destination station of the first journey (J1.destination) and the origin station of the second journey (J2.origin) in the journey pairs are selected. If selected stations match or are in close proximity (walking distance,  $\leq 500$  m), the *stay-time* is extracted using the origin time of the second journey (J2.origin time) and destination time of the first journey (J1.destination time). The results selected are those which pass a predefined threshold for *stay-time*. The next stage is to pass these selected stations through the criteria of *visit-frequency* threshold. If a station is identified more than the defined threshold for *visit-frequency*, it is classified as a work location for that user. Based on this criterion, users can have one or more work locations. Similarly, it is also possible that the algorithm fails to highlight any station as a work location if no location meets the expected criteria (Figure 3).

In the situation where more than one station appears as a work station, Alexander et al. (2015) and Wang et al. (2017) applied criteria based on outcome of frequency and distance from home location to identify the work location. The limitation however of such an approach is that it fails to take

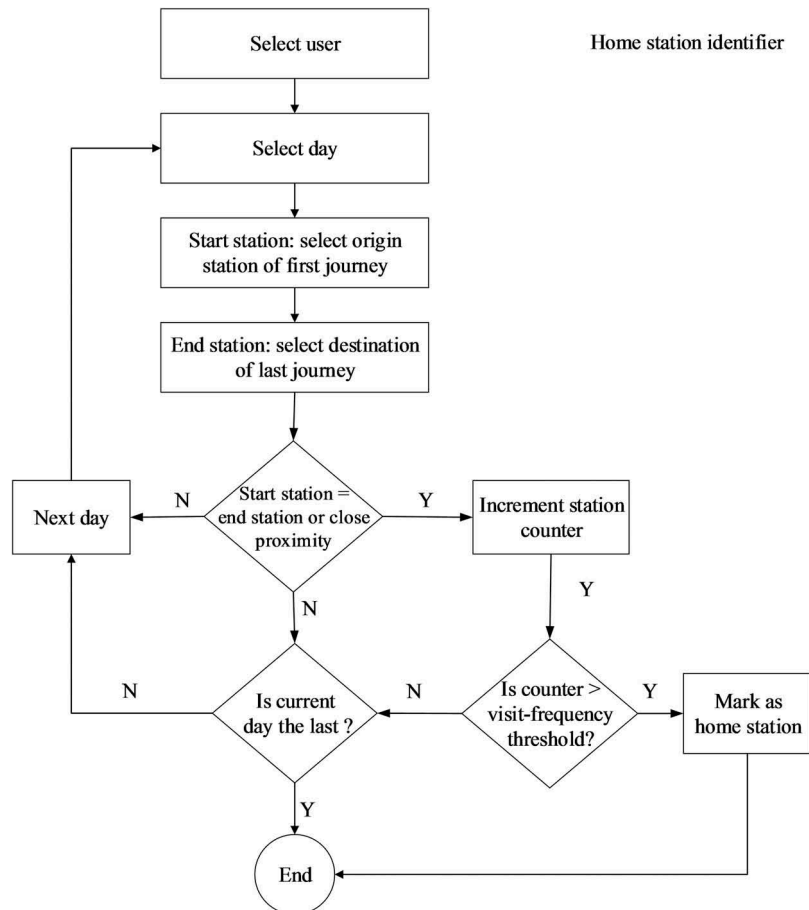


Figure 2. Flowchart of home location identifiers (algorithm 1) .

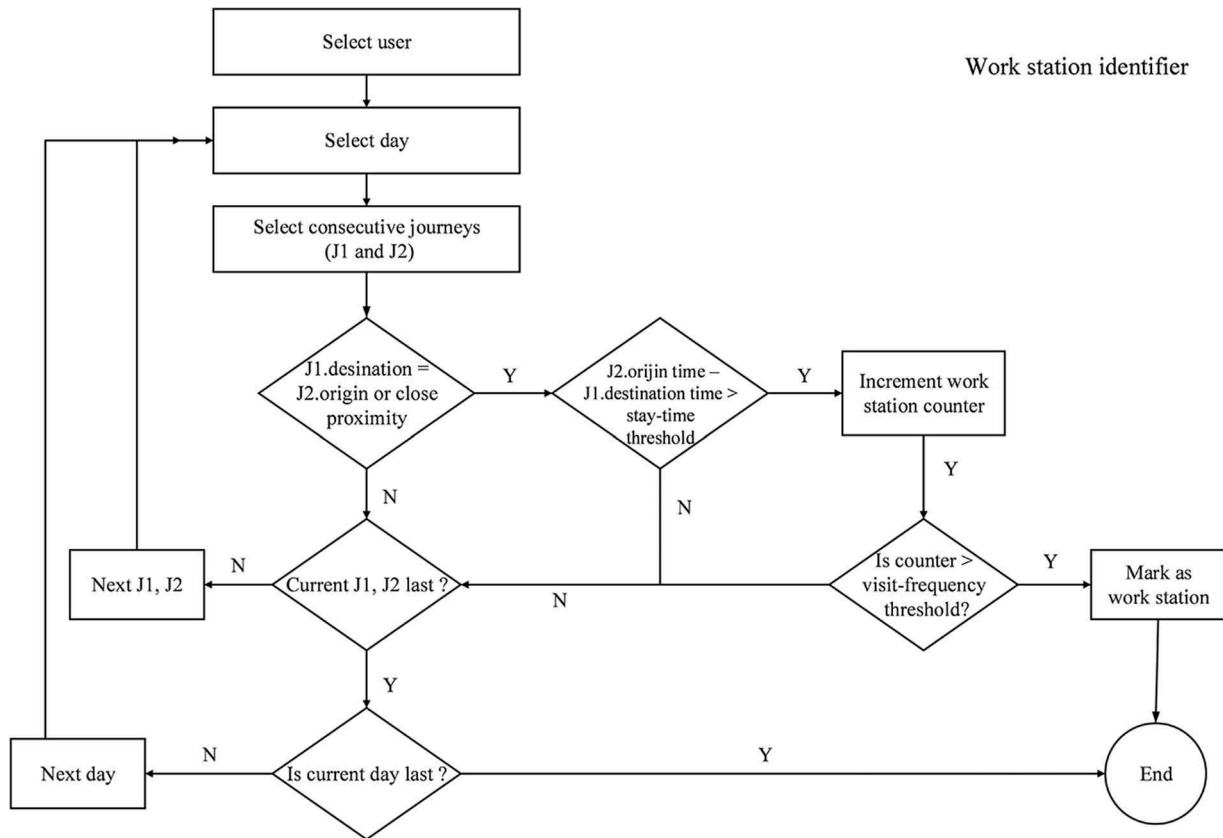


Figure 3. Flowchart of work location identifiers (algorithm 2).

account of the duration of the work activity, due to which multiple candidates for work locations are identified. Additionally, limiting the work location to a single station can be inaccurate as an individual might have more than one station close to their work. Therefore, an approach based on distance (walking distance,  $\leq 500$  m), presented by rank, using frequency attributes and *stay-time* duration to provide a meaningful outcome.

The above work location identification algorithm is based on the *stay-time* and *visit-frequency* of consecutive journeys. We are using *stay-time* criterion based on the pragmatic assumptions made by Devillaine, Munizaga, and Trépanier (2012) to identify work locations by using activity duration, and we are following Hasan et al.'s (2012) example by using *visit-frequency* as an indicator in a rule-based algorithm for the identification of work locations. The combination of both indicators along with the close proximity distance forms the basis of work location identification used in this study.

### 3.4. Validation process

The identification of primary locations is significant to understand journey purpose. Three separate validation methods have been applied in this work to gauge the accuracy of the results.

The first method applied to use the information in the LTDS data to validate the results of the heuristic method at the level of postcode district. The approach to validation applied in this section matches the home and work stations identified to the postcode district of the station. This information is then compared with the LTDS data. To validate the results, the Location Weighted Average (LWA) has been calculated using the correctly identified user locations as a percentage of total user locations in the postcode district and the total number of correct locations in the dataset as described in the following:

$$LWA = \left[ \frac{VR}{TVR} \times \frac{TVR}{TU_{in\ postcode\ district}} \right] \times 100 \quad (1)$$

where VR is the validated results, TVR is the total validated results, and TU is total number of users in the LTDS dataset.

The second approach of the validation process compares the results to another model described by Hasan et al. (2012). The model uses the distribution of people's most-visited places as the key driver to classify work and home locations in London. To achieve this, the validation process looks at a subset of user data for which we have identified locations through both methods.

Finally, LTDS dataset available for the evaluation in this study was based on one-day data in a typical year. In a dynamic city like London where a large

number of people move in and out of the city, as well as people who change places of work and residence, this makes the old survey data inaccurate for evaluation. Therefore, a labeled dataset was created and an evaluation was carried out using a recent survey via interviews and by collecting travel history data.

## 4. London case study

### 4.1. Study area and data description

London is one of the most populated and fast-growing cities in the world. It also has one of the oldest and most comprehensive public transports in the form of the London Underground, also known as the Tube, which began in 1863 and now comprises 270 stations covering over 400 km.

The aim of the case study was to apply a proposed model on smart card data for Transport for London (TfL). The card is valid on all London public transport systems such as London Underground, the bus network, the Docklands Light Railway, London Overground, Tramline, some river boat services, and most National Rail services within the London Fare Zones. The framework proposed in this study is able to capture intermodal commuting patterns (train–bus, train–bike) as long as the complete OD information is captured in the journey record.

*Oyster Data:* Oyster is a smart card used to hold travel credit and travel passes for journeys carried out on the TfL network. With the help of Oyster cards, TfL is able to keep a record of individual journeys carried out using the card. The volume of journeys carried out using the Oyster cards on the TfL network is more than 80% of the 3 million journeys carried out each day on the network (TfL 2016). Ninety-five percent of all Oyster card usage is for London's underground and bus journeys although the Oyster card is used on multiple modes of transportation across London (Gordon 2012). One of the limitations of the TfL dataset is the incomplete journey information captured for bus journeys. As TfL systems do not currently capture the alighting information from its bus journeys, these journeys are excluded from the analysis. Such journeys, however, can be included with an improvement to the model, where missing alighting information is identified as a sub-step within the process. In this study, the sample data available consisted of a total of 60 million journeys. Since the processing of such large amounts of data is resource intensive, smart card records of 10,000 TfL users were randomly selected for further analysis. After excluding users with only single journeys, the data of 9900 individuals with a total of 1,823,906 complete journey records were considered for further analysis. This data covered the period of October–November 2013. These data were prepared for each

individual, extracting details of their daily movements on the TfL network (excluding bus and tram journeys) such as date, entry time and station, exit time, station, and transport mode.

TfL Oyster card data provided an important source of information for this study, but it was not adequate for the verification of the model. To validate the results of the heuristic primary location model, labeled data were required. As part of this study, Oyster card travel records of 25 volunteers were collated, along with their home and work locations, since TfL users can access their personal data including travel records. Data such as date, entry time, entry station, exit time, exit station, and transport mode were preprocessed to extract daily movements on the TfL network. It consisted of a total 6243 journey records for two months (excluding bus and tram journeys). These volunteer survey data were used to validate the proposed model and potentially eliminate the need for expensive travel demand surveys

*LTDS data:* The LTDS data are a single day recorded surveys around 8000 households in a typical year. It comprises a detailed household questionnaire focused on travel in Greater London. These questions provide insight into the socio-demographics of household members and the mode, travel time, travel purpose, origin, and destination of each journey stage. Clearly, the LTDS does not offer adequate detail on passenger behavior across the entire network. However, travel patterns derived from Oyster data are still comparable to LTDS (Seaborn, Attanucci, and Wilson 2009).

Part of the LTDS dataset includes the Oyster card ID numbers of the respondents. This enabled TfL to match the information in the questionnaire to the actual journey records of those people. These linked journeys provide an invaluable source of test data to validate the model proposed in this study. The data were a pool of 5,718,644 records for a total of 10,895 unique users for the period 2011–2014. After excluding the single-journey users from linked journeys, a total of 369,745 journeys for 9479 users were selected for the months of January–March 2014. From the selected records, bus journeys were also removed to arrive at a total journey count of 124,031 eligible for further analysis.

### 4.2. Results of the case study

Regular commuters use the network for their daily activities that involve travel to and from their primary locations. After the preprocessing of the dataset, the regularity of usage for individual users is examined through journey count. To create a meaningful dataset, it is necessary to make assumptions to define acceptable journey count threshold. Since a minimum of 2 journeys are required to carry out an activity, a

minimum of 2 and maximum of 60 journey counts are considered to examine the regularity of usage.

Figure 4 highlights that the fall in the number of regular users plateaus above a journey count of 10, whereas between the thresholds of 2 and 10, the reduction in the number of regular commuters is approximately 52%. Therefore, a threshold of 10 is considered appropriate to define the regularity of usage.

To reveal the temporal regularity of peoples' mobility, *visit-frequency* was investigated. To attain a level of confidence in the results, the heuristic primary location model was presented against the number of different *visit-frequency* counts as a measure of mobility patterns. Different indicator values provided a different outcome for the algorithms based on the available data.

Figure 5 illustrates the number of individuals captured home and work locations using the different visit frequency values. It can be seen that a *visit-frequency* value of 5 provides a demarcation point. The number of individuals decreases at a slower rate above the value of 5 as compared to below the value of 5. Therefore, *visit-frequency* 5 is applied in the heuristic primary location model. A *visit-frequency* threshold value of 5 means the algorithm must correctly identify the home location five times before it can be classified in the outcome as a home location. The same threshold value is applied to the work locations.

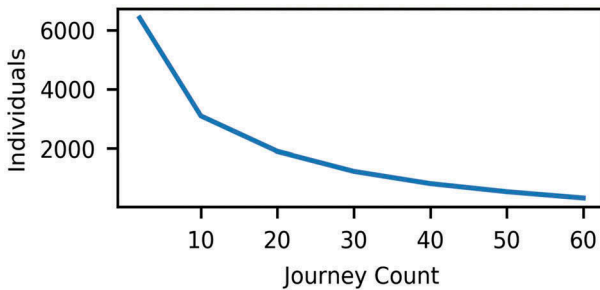


Figure 4. Presents the number of unique users that have carried out journeys within different values of journey count threshold during the selected period.

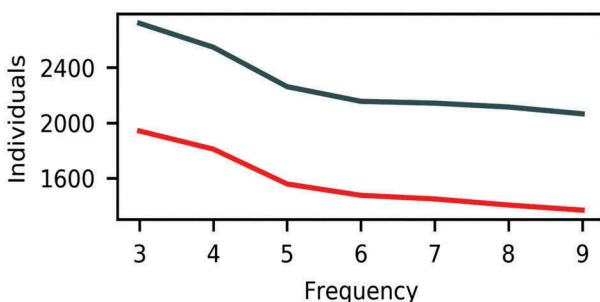


Figure 5. The figure highlights the impact of *visit-frequency* indicator in the identification of an individual's home (blue line) and work (red line) locations.

Another important indicator with regard to the temporal patterns is the *stay-time* duration describing activities between locations gathered from smart card data. It identifies the *stay-time* between consecutive journeys and enables the identification of the work location.

Figure 6 highlights the impact of *stay-time* in the identification of work location for the dataset within a range of 2–14 h. As the *stay-time* duration increases, the number of users identified decreases. There is a decline of 14% for the *stay-time* duration of up to 4 h. The drop in the number of the users identified is less significant between 4 and 8 h as compared to the first 4 h. It is expected that the activities durations of around 8 h would cover most of the regular commuters. Moreover, the number of users identified with *stay-time* duration greater than 8 h sees a significant drop at almost 60%. To capture the full-time as well as part-time workers, greater than 4 h was considered as an appropriate *stay-time* duration for this study.

The heuristic primary location model applied using the indicators of *visit-frequency* and *stay-time* duration was able to identify home and work locations. The results are presented in Figures 7 and 8 aggregate data at the level of stations. The points on the map represent the number of users that identify a given station as their home or work location.

Figure 7 demonstrates the well-connected nature of the London transportation network with home locations dispersed evenly around London. Top two home stations highlighted from the analysis are Brixton and Stratford. It also shows that the outer boroughs of London were represented by smaller data points in comparison to the inner London stations. This is representative of the high population density in central London.

As can be seen in Figure 8, the key work locations are identified, particularly the centers of financial services around the City of London. Locations outside of central London were also identified as the work locations, such as Ealing Broadway, Stratford, and Brixton. These locations are examples of commercial centers outside central London.

The contribution of the Figures 7 and 8 is that it highlights the locations such as Stratford, Brixton,

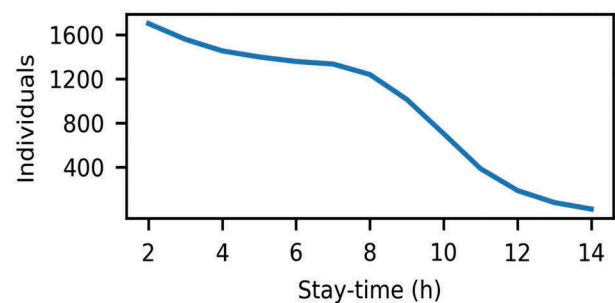
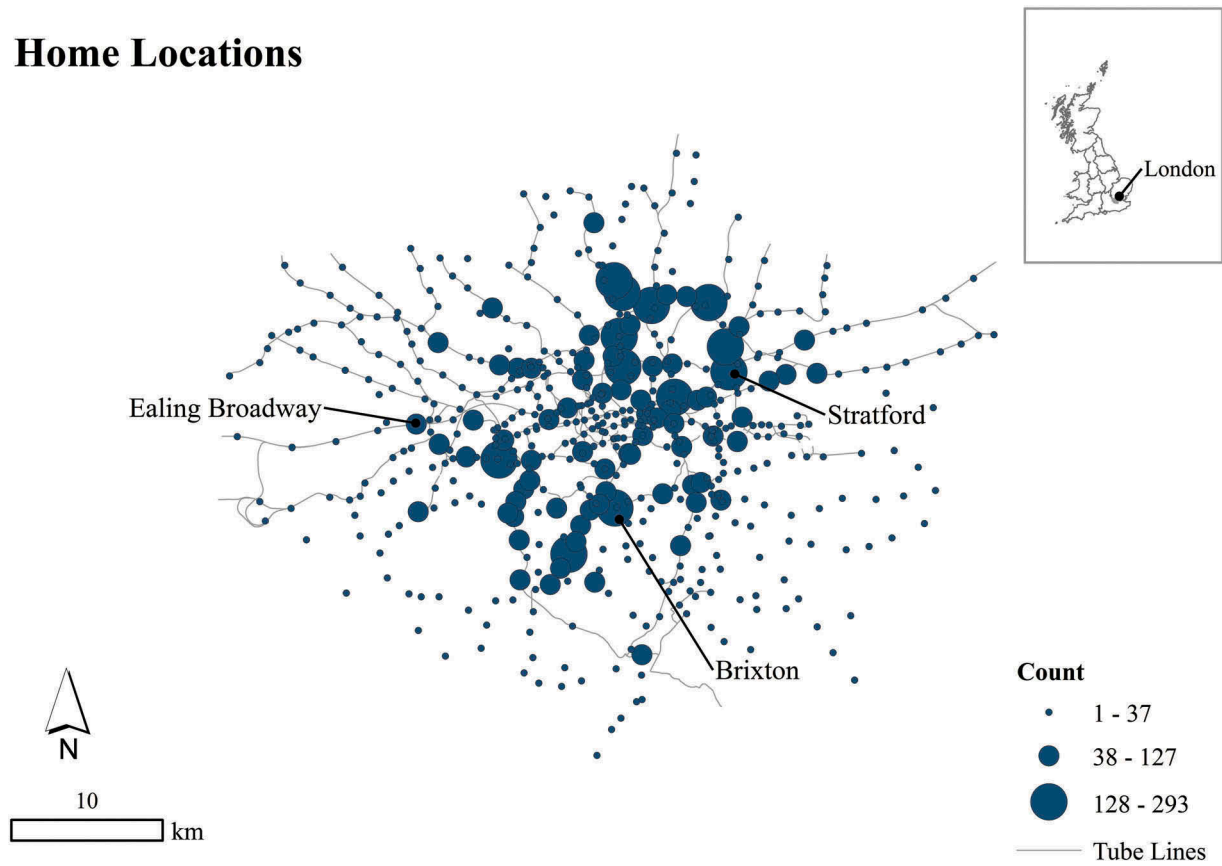


Figure 6. The number of individuals identified at work location using different *stay-time* duration values.

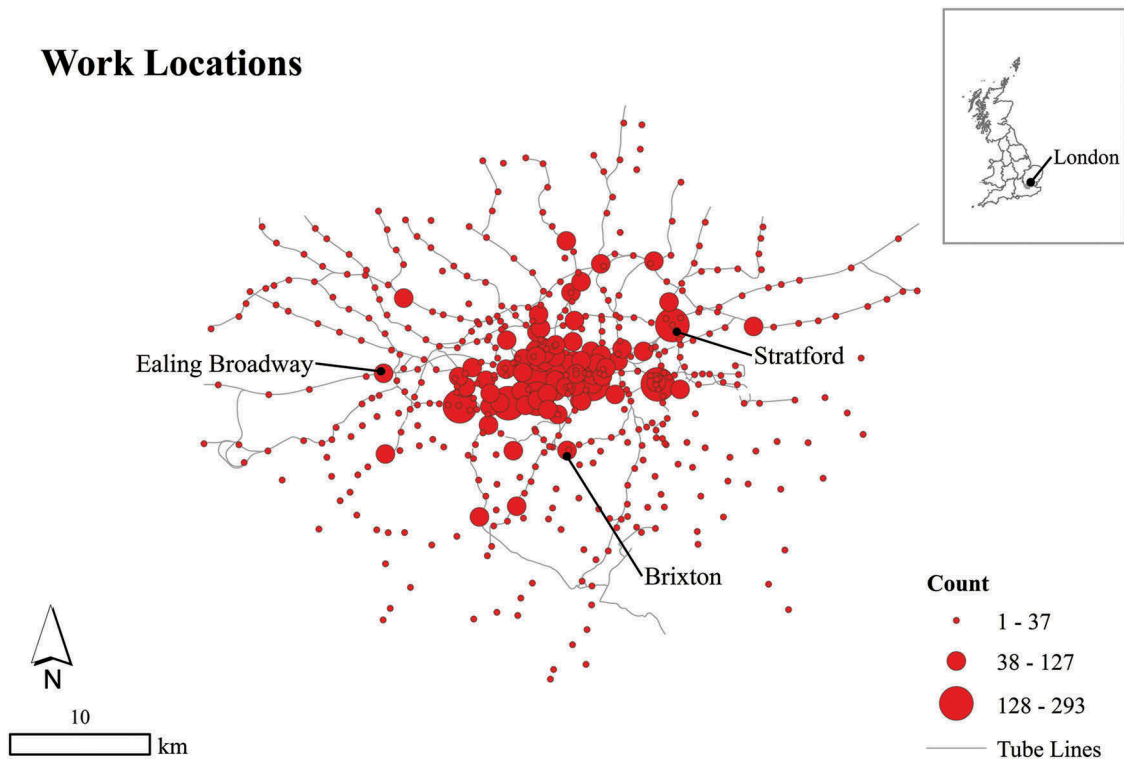


## Home Locations



**Figure 7.** Results of algorithm 1 presenting home locations around train/tube stations in London. The data on the map are aggregated at the level of the station. Each data point represents the number of users that identified a given station as their home location.

## Work Locations



**Figure 8.** Results of algorithm 2 presenting work locations around train/tube stations in London. The data on the map are aggregated at the level of the station. Each data point represents the number of users that identified a given station as their work location.

and Ealing Broadway that are significant as both the home and work locations. In contrast to Long, Zhang, and Cui (2012) and Wang et al. (2017), these figures illustrate that large cities are often not clearly segregated by land use; the residential areas frequently contain work locations and other land-use types.

#### 4.3. Validating the results of the model

Validation of the analysis was carried out using the matched-journeys data at the postcode district level. It was a smaller dataset compared to the available smart

card data. Matched-journeys data captured information about the home and work locations of individuals. This made the LTDS data invaluable as it provided a source of validation for the algorithm applied in this study.

Figure 9 presents the result of the algorithm for the identification of home and work locations aggregated at the level of each London postcode district. It illustrates that some of the stations in London can be considered home locations as well as work locations.

The weighted outcomes were calculated using the correctly identified user locations as a percentage of total user locations in the postcode district and the total number of correct locations in the dataset. The

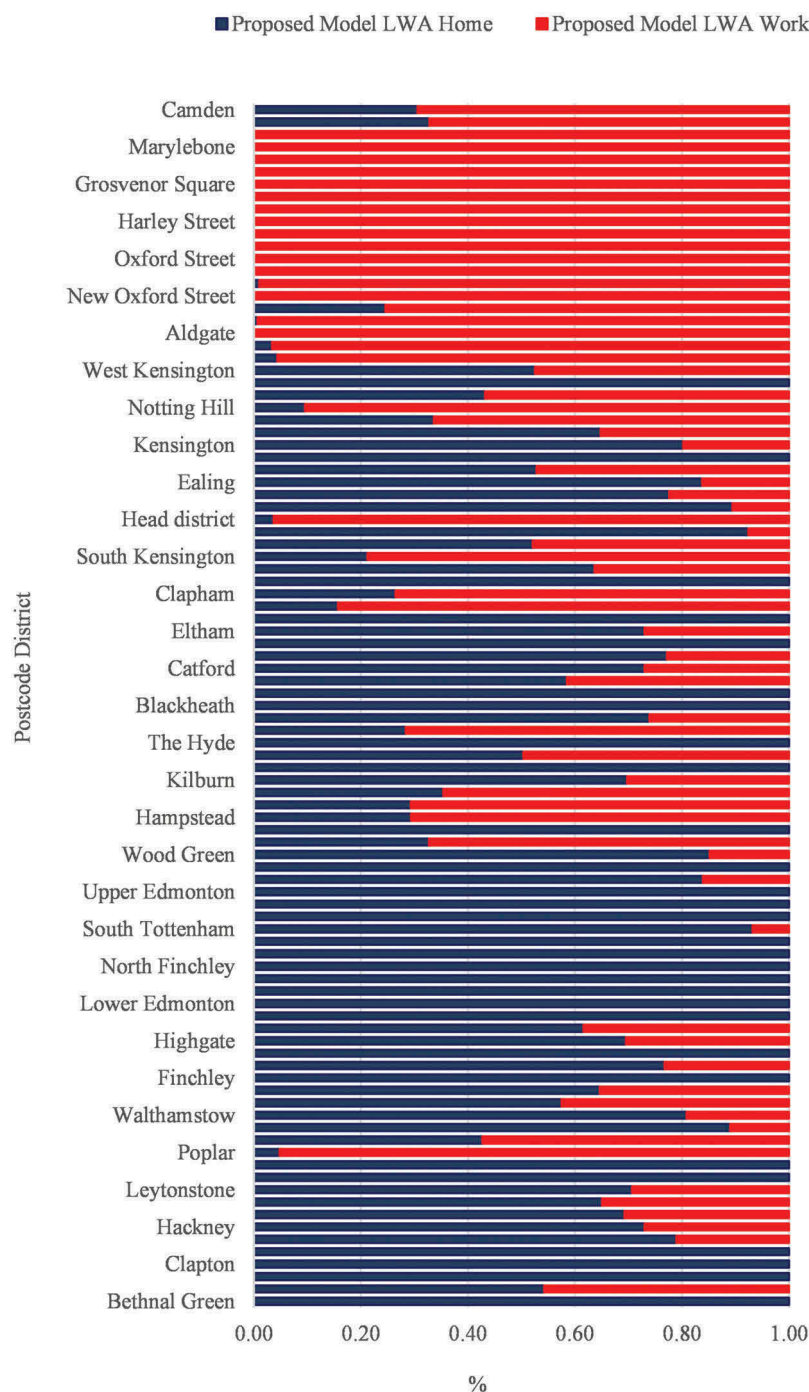


Figure 9. Proposed model LWA home and work location validation results.

results of the proposed model were compared to the LTDS dataset; the algorithm identified 82% of the home users with the same location as the LTDS data. For work locations, the percentage of users verified accurately using LTDS data was 60%.

The second part of the validation demonstrates a comparative analysis of another rule-based approach in London (Hasan et al. 2012) as the benchmark. The existing model results when validated against the LTDS demographic dataset corresponded to a success rate of 59% for home location and 35% for work locations (excluding missing LTDS data). As a result, the heuristic primary location model in this study demonstrates a better identification of home and work locations as compared to more simplistic methods.

The proposed model provided better accuracy than the existing model as a benchmark presented for London, but accuracy for home and work locations in London's highly changeable context is difficult due to the city's high rate of residential and employment flux. This is especially the case for the LTDS dataset, and is thus only an approximate gauge for comparisons. Because such survey data is limited in its ability to provide an accurate source for evaluation, we have supplemented the work with additional verification against a more recent dataset of journeys and locations collected from volunteer users. When the same method was applied to this new dataset, the accuracy of the home and work locations was identified at 97% and 93%, respectively.

## 5. Conclusions

This study aimed to develop a framework for inferring urban mobility from smart card data by developing a heuristic primary location model. The model uses journey counts as an indicator of usage regularity and *visit-frequency* to identify activity locations for regular commuters, and *stay-time* for the classification of work and home locations and activities. London is taken as a case study, and the model results were validated against data from the LTDS and volunteer survey data. Results demonstrate that the proposed model is able to detect meaningful home and work places with high precision at 97% and 93%, respectively, from labeled smart card dataset. Thus, we conclude that a combination of limited labeled data collected by means of representative user surveys and a large amount of unlabeled data from the Oyster network together has the potential to vastly improve the way mobility analysis is carried out in large cities.

It can be concluded from this study that smart card data presents substantial prospects for the understanding of commuter behavior and can provide an accurate and more reliable picture compared to user mobility profiles generated using sample

surveys. This approach to human mobility can, in turn, improve understandings of wider mobility patterns at an aggregate level. It can help city officials recognize the complex underlying factors of transportation use and develop efficient and sustainable urban transportation systems.

This study has a number of limitations to be addressed in future work. First, bus journeys were excluded because of a lack of alighting point information. These journeys can be included with an enhancement of the model, where the missing information is also identified as a sub-step within the identification process. Second, distances covered by walking cannot be calculated with the smart card data to exactly identify a home or work location, nevertheless GPS data from mobile phones can be used in this context. Finally, an interesting topic to explore is the identification of user locations and activities that are less clearly defined as being home and work locations. The rules concerning these are more complex and would require a more dynamic approach to identification.

## Funding

This work was funded by the Economic and Social Research Council (ESRC) in the United Kingdom [grant number 1477365].

## Notes on contributors

*Nilufer Sari Aslam* is a PhD student in SpacetimeLab at UCL. Her research interests are transportation networks and data mining techniques. As part of her PhD study, she is currently investigating the user mobility from spatial-temporal big datasets.

*Tao Cheng* is currently a professor in GeoInformatics at UCL. She is also the director of SpaceTimeLab for Big Data Analytics. Her current research interests include space-time analytics and big data mining with applications in transport modeling.

*James Cheshire* is a senior lecturer in Quantitative Geography at UCL. His research focuses on the use of big datasets for the study of social science.

## ORCID

Nilufer Sari Aslam  <http://orcid.org/0000-0003-4552-5989>

Tao Cheng  <http://orcid.org/0000-0002-5503-9813>

James Cheshire  <http://orcid.org/0000-0003-4552-5989>

## References

- Agard, B., C. Morency, and M. Trépanier. 2007. "Mining Public Transport User Behaviour from Smart Card Data." *IFAC Proceedings Volumes* 39 (3): 399–404. doi:10.3182/20060517-3-FR-2903.00211.
- Ahas, R., S. Silm, O. Jarv, E. Saluveer, and M. Tiru. 2010. "Using Mobile Positioning Data to Model Locations

- Meaningful to Users of Mobile Phones.” *Journal of Urban Technology* 17 (1): 3–27. doi:10.1080/10630731003597306.
- Alex, A. P., M. V. Saraswathy, and K. P. Isaac. 2016. “Modelling of Daily Activity Schedule of Workers Using Unsupervised Machine Learning Technique.” *International Journal for Traffic and Transport Engineering* 6 (1): 77–91. doi:10.7708/2217-544X.
- Alexander, L., S. Jiang, M. Murga, and M. C. González. 2015. “Origin-Destination Trips by Purpose and Time of Day Inferred from Mobile Phone Data.” *Transportation Research Part C: Emerging Technologies* 58: 240–250. doi:10.1016/j.trc.2015.02.018.
- Alsger, A., T. Ahmad, M. Mahmoud, F. Luis, and H. Mark. 2016. “Public Transport Origin-Destination Estimation Using Smart Card Fare Data.” *Sante Publique* 28 (3): 391–397.
- Anda, C., A. Erath, and P. J. Fourie. 2017. “Transport Modelling in the Age of Big Data.” *International Journal of Urban Sciences* 21 (Sup1): 19–42. doi:10.1080/12265934.2017.1281150.
- Bagchi, M., and P. R. White. 2005. “The Potential of Public Transport Smart Card Data.” *Transport Policy* 12 (5): 464–474. doi:10.1016/j.tranpol.2005.06.008.
- Barry, J., R. Newhouser, A. Rahbee, and S. Sayeda. 2002. “Origin and Destination Estimation in New York City with Automated Fare System Data.” *Transportation Research Record: Journal of the Transportation Research Board* 1817 (02): 183–187. doi:10.3141/1817-24.
- Chakirov, A., and A. Erath. 2012. “Activity Identification and Primary Location Modelling Based on Smart Card Payment Data for Public Transport Smart Card Payment Data for Public Transport.” Paper presented at the 13th International Conference on Travel Behaviour Research, Toronto, July 15–20.
- Deng, Z., and M. Ji. 2010. “Deriving Rules for Trip Purpose Identification from GPS Travel Survey Data and Land Use Data: A Machine Learning Approach.” Paper presented at the 7th International Conference on Traffic and Transportation Studies, Kunming, China, August 3–5.
- Devillaine, F., M. Munizaga, and M. Trépanier. 2012. “Detection of Activities of Public Transport Users by Analyzing Smart Card Data.” *Transportation Research Record: Journal of the Transportation Research Board* 2276 (3): 48–55. doi:10.3141/2276-06.
- Gordon, J. B. 2012. Intermodal Passenger Flows on London’s Public Transport Network. Accessed 16 April 2017. <https://dspace.mit.edu/bitstream/handle/1721.1/78242/830539087-MIT.pdf?sequence=2>
- Han, G., and K. Sohn. 2016. “Activity Imputation for Trip-Chains Elicited from Smart-Card Data Using a Continuous Hidden Markov Model.” *Transportation Research Part B* 83: 121–135. doi:10.1016/j.trb.2015.11.015.
- Hasan, S., C. M. Schneider, S. V. Ukkusuri, and M. C. González. 2012. “Spatiotemporal Patterns of Urban Human Mobility.” *Journal of Statistical Physics* 151: 304–318. doi:10.1007/s10955-012-0645-0.
- Hasan, S., X. Zhan, and S. V. Ukkusuri. 2013. “Understanding Urban Human Activity and Mobility Patterns Using Large-Scale Location-Based Data from Online Social Media.” Paper presented at the 2nd ACM SIGKDD International Workshop on Urban Computing – UrbComp’13, Chicago Sheraton, Chicago, USA, August 11. <http://dl.acm.org/citation.cfm?id=2505821.2505823>.
- Isaacman, S., R. Becker, R. Cáceres, S. Kobourov, M. Martonosi, and J. Rowland. 2011. “Identifying Important Places in People’s Lives from Cellular Network Data 1 Introduction.” *Pervasive Computing* 6696: 133–151.
- Jiang, S., Y. Yang, S. Gupta, D. Veneziano, S. Athavale, and M. C. González. 2016. “The TimeGeo Modeling Framework for Urban Motility without Travel Surveys.” *Proceedings of the National Academy of Sciences of the United States of America* 113 (37): E5370–E5378. doi:10.1073/pnas.1524261113.
- Li, G., L. Yu, W. S. Ng, W. Wu, and T. G. Shen. 2015. “Predicting Home and Work Locations Using Public Transport Smart Card Data by Spectral Analysis.” Paper presented at the 18th IEEE Conference on Intelligent Transportation Systems, Las Palmas de Gran Canaria, Spain, September 15–18.
- Long, Y., Y. Zhang, and C. Cui. 2012. “Identifying Commuting Pattern of Beijing Using Bus Smart Card Data.” *Acta Geographica Sinica* 67 (10): 1339–1352.
- Lotero, L., R. G. Hurtado, L. M. Floria, and J. Gómez-Gardeñes. 2016. “Rich Do Not Rise Early: Spatio-Temporal Patterns in the Mobility Networks of Different Socio-Economic Classes.” *Royal Society Open Science* 3 (10): 1–12. doi:10.1098/rsos.160131.
- Luong, T. B. T. 2015. “Human Activity Recognition: A Data-Driven Approach.” UC Irvine Electronic Theses and Dissertations, Accessed 18 April 2017. <https://escholarship.org/uc/item/4w98w1zd>.
- Maat, K., B. van Wee, and D. Stead. 2005. “Land Use and Travel Behaviour: Expected Effects from the Perspective of Utility Theory and Activity-Based Theories.” *Environment and Planning B: Planning and Design* 32 (1): 33–46. doi:10.1068/b31106.
- Manley, E., C. Zhong, and M. Batty. 2018. “Spatiotemporal Variation in Travel Regularity through Transit User Profiling.” *Transportation* 45 (3): 703–732. doi:10.1007/s11116-016-9747-x.
- Pelletier, M.-P., M. Trépanier, and C. Morency. 2011. “Smart Card Data Use in Public Transit: A Literature Review.” *Transportation Research Part C: Emerging Technologies* 19 (4): 557–568. doi:10.1016/j.trc.2010.12.003.
- Seaborn, C., J. Attanucci, and N. H. M. Wilson. 2009. “Analyzing Multimodal Public Transport Journeys in London with Smart Card Fare Payment Data.” *Transportation Research Record: Journal of the Transportation Research Board* 2121 (1): 55–62. doi:10.3141/2121-06.
- TfL. 2016. “Oyster Card.” Accessed 22 April 2017. <https://tfl.gov.uk/corporate/publications-and-reports/oyster-card>.
- Trépanier, M., N. Tranchant, and R. Chapleau. 2007. “Individual Trip Destination Estimation in a Transit Smart Card Automated Fare Collection System.” *Journal of Intelligent Transportation Systems* 11 (1): 1–14. doi:10.1080/15472450601122256.
- Wang, Y., G. H. de Almeida Correia, E. de Romph, and H. J. P. Timmermans. 2017. “Using Metro Smart Card Data to Model Location Choice of After-Work Activities: An Application to Shanghai.” *Journal of Transport Geography* 63: 40–47. doi:10.1016/j.jtrangeo.2017.06.010.
- Zhao, J., A. Rahbee, and N. H. M. Wilson. 2007. “Estimating a Rail Passenger Trip Origin–Destination Matrix Using Automatic Data Collection Systems.” *Computer-Aided Civil and Infrastructure Engineering* 22 (5): 376–387. doi:10.1111/j.1467-8667.2007.00494.x.
- Zou, Q., X. Yao, H. Wei, and H. Ren. 2016. “Detecting Home Location and Trip Purposes for Cardholders by Mining Smart Card Transaction Data in Beijing Subway.” *Transportation* 45 (3): 919–944. doi:10.1007/s11116-016-9756-9.