

Semantic enrichment of secondary activities using smart card data and point of interests: a case study in London

Nilufer Sari Aslam , Di Zhu , Tao Cheng , Mohamed R. Ibrahim & Yang Zhang

To cite this article: Nilufer Sari Aslam , Di Zhu , Tao Cheng , Mohamed R. Ibrahim & Yang Zhang (2020): Semantic enrichment of secondary activities using smart card data and point of interests: a case study in London, Annals of GIS, DOI: [10.1080/19475683.2020.1783359](https://doi.org/10.1080/19475683.2020.1783359)

To link to this article: <https://doi.org/10.1080/19475683.2020.1783359>




© 2020 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group, on behalf of Nanjing Normal University.



Published online: 01 Aug 2020.



[Submit your article to this journal](#) 



[View related articles](#) 



[View Crossmark data](#) 

Semantic enrichment of secondary activities using smart card data and point of interests: a case study in London

Nilufer Sari Aslam ^a, Di Zhu ^{a,b}, Tao Cheng ^a, Mohamed R. Ibrahim ^a and Yang Zhang ^a

^aSpaceTimeLab, Department of Civil, Environmental and Geomatic Engineering, University College London (UCL), London, UK; ^bInstitute of Remote Sensing and Geographical Information Systems, School of Earth and Space Sciences, Peking University, Beijing, China

ABSTRACT

The large volume of data automatically collected by smart card fare systems offers a rich source of information regarding daily human activities with a high resolution of spatial and temporal representation. This provides an opportunity for aiding transport planners and policy-makers to plan transport systems and cities more responsively. However, there are currently limitations when it comes to understanding the secondary activities of individual commuters. Accordingly, in this paper, we propose a framework to detect and infer secondary activities from individuals' daily travel patterns from the smart card data and reduce the use of conventional surveys. First, we proposed a 'heuristic secondary activity identification algorithm', which uses commuters' primary locations (home & work) and the direction (from & to) information to identify secondary activities for individuals. The algorithm provides a high-level classification of the activity types as before-work, midday and after-work activity patterns of individuals. Second, this classification is semantically enriched using Points of Interests to provide meaningful insights into individuals' travel purposes and mobility in an urban environment. Lastly, using the transit data of London as a case study, the model is compared with a volunteer survey to demonstrate its effectiveness and offering a cost-effective method to travel demand research.

ARTICLE HISTORY

Received 28 October 2019
Accepted 10 June 2020

KEYWORDS

Activity identification; smart card data; heuristic (rule-based) modelling; secondary activities; trip purpose; urban activity pattern

1. Introduction

'Activities' are vital to understanding travel behaviour, mobility patterns and traffic volumes in an urban setting (Ben-akiva, Bowman, and Gopinath 1996). Activities that occur at home and work/school (for adults and students, respectively) locations are referred as primary activities, and the rest of the activities (such as eating, shopping, entertaining, etc.) are referred as secondary activities. As people have more spare time and surplus disposable income, time spent in secondary activities is taking a larger percentage in their daily routines (Lu and Gu 2011; Zhong et al. 2014). Therefore, the need to analyse these activities is more important than ever.

The identification of these secondary activities is not only beneficial to transport planners for a better appreciation of individuals' travel behaviour but also for commercial organizations in the context of consumer behaviour (Goulet-Langlois 2016) and for economists, providing a useful insight into the quality of life and aspiration (Nakamura et al., 2016). This expands the scope of research from urban transportation in travel behaviour and mobility (Yang et al. 2019), trip purposes (Alsger et al., 2018), accessibility (Saif, Zefreh, and Torok 2019) to social studies (Zhu et al. 2017)

Secondary activities were investigated in the literature using activity-travel demand models derived from conventional travel surveys (Arentze and Timmermans 2007; Rasouli and Timmermans 2015). A relatively small sample size (only a one-day travel diary) was used to estimate travel demand for the whole population. Such surveys are expensive and time-consuming. Gathering smart data is significantly more efficient in terms of both cost and time due to the automated nature of these systems (Pelletier, Trépanier, and Morency 2011). In addition, they are usually available for a much larger population and for a longer period, which assists in understanding mobility behaviours and travel flows. (Bagchi and White 2005). Notwithstanding the wide range of positive characteristics, smart card data present several challenges such as: estimating a commuter's destination if public transport does not ask for alighting information (Gordon et al. 2013), making demographic predictions if socio-demographic information is not accessible due to privacy concerns (Zhang, Cheng, and Sari Aslam 2019; Zhang and Cheng 2020), detecting activities in order to estimate a trip's purpose by linking smart card data with auxiliary data sources, such as land use maps and POIs (Devillaine, Munizaga, and Trépanier 2012; Kuhlman 2015; Sari Aslam and Cheng 2018; Yang et al. 2019).

CONTACT Nilufer Sari Aslam  n.aslam.11@ucl.ac.uk

The advantage of using smart card data is their ability to reveal an individual's spatial-temporal activity pattern as a sequence of activity locations and durations on a daily basis. Mining activity patterns from smart card data is crucial to having an accurate estimate of travel purpose (Ma et al. 2013, 2017). However, activity identification and inferring models using smart card data rarely implemented on public transport net works, even in cases where they have been applied, the scope of the models has been limited to primary activities (Chu and Chapleau 2010; Chakirov and Erath 2012; Devillaine, Munizaga, and Trépanier 2012; Zou et al., 2016; Yang et al. 2019) except for Wang et al. (2017) and Alsger et al. (2018). According to Wang et al. (2017), after-work activities were defined using time constrains once home and work locations had been identified within a sequence of activity patterns. However, because the sequence of activities differs for each person daily, extracting them with temporal attributes only may overlook some of the secondary activities. Alternatively, Alsger et al. (2018) has incorporated some of secondary activities using a wide array of auxiliary data sources such as O-D survey, land-use data, household travel surveys and weather reports in a rule-based approach. Although Alsger et al. (2018) shed light on inferring some of secondary activities such as shopping and recreation activities, the method is not applicable to large smart card data when the detailed surveys and auxiliary attributes are not available.

With this in mind, the motivation of this study is two-fold. First, the study proposes an algorithm to identify, on a daily basis, the secondary activities within a sequence of activity patterns from smart card data in order to obtain the dynamic of individual mobility – 'where and when individuals move within the city'. Second, the study proposes a feasible framework to be able to infer activity types and travel purposes in order to understand 'why individuals move within the city.' Hence, the framework includes following: first, we extract individual activities using spatial (distance measure) and temporal (transfer time) information of smart card data. We then identify primary locations (home & work) using boarding and alighting stops, activity duration, and the frequency characteristic of the smart card data (Sari Aslam, Cheng, and Cheshire 2019). After that, secondary activities are detected based on their direction (from & to) relative to the primary locations, classified into four types, i.e., before-work, midday, after-work and undefined activities to represent individuals' activity patterns (Pinjari et al., 2007; Rasouli and Timmermans 2015; Wang et al. 2017). This classification is considered to provide a more meaningful inference of activity and travel demand (Pinjari et al. 2007; Ma et al. 2013; Wang

et al. 2017). Finally, using POIs to enrich the data we classify activities into one of the semantic sub-categories : eating, entertainment, shopping, work, other, to represent an individual trip's purposes.

The contribution of this study is four-fold:

- extracting individual activities from smart card data and using a heuristic activity identification algorithm to reveal secondary activities.
- supporting trip chaining models using the location of activities from smart card data only.
- investigating individual mobility using a large individual-travel dataset, i.e. smart card data, as an alternative to the expensive travel demand survey.
- combining both the O-D information and the tube stations' socio-functional information from POIs to enrich station-based secondary activities (the current state-of-the-art approach).

The next section of the paper is to details the research framework and methods, which is followed by the case study in London. The last section summarizes the conclusions and future directions of the work.

2. Method

This study focuses on the semantic meaning of secondary activities identified as part of tube/train commutes around home and work locations from smart card data. Primary activities provide people's regular move within cities (Wang, Homem, et al. 2017) combined with the nearest spatial information such as 'where you come from before an activity' and 'where you go to after an activity' to estimate secondary activities from smart card data.

In our study, secondary activities and their semantic significance are inferred via the four steps presented in Figure 1. First, trips and activities are extracted from smart card data. Then we detect secondary activities in two steps. The first step is to determine the home and work locations as the anchor points for each user (Sari Aslam, Cheng, and Cheshire 2019). The second step is to combine the primary locations and direction (from & to) information and create the 'what-if' scenarios to allow us to classify secondary activities into the four empirical types: before-work activity, midday activity, after-work activity and undefined activity. Finally, secondary activities are further categorized using auxiliary information such as the POIs near each secondary activities' location. Hence, an individual trip's purpose is investigated where, when, and why people spent

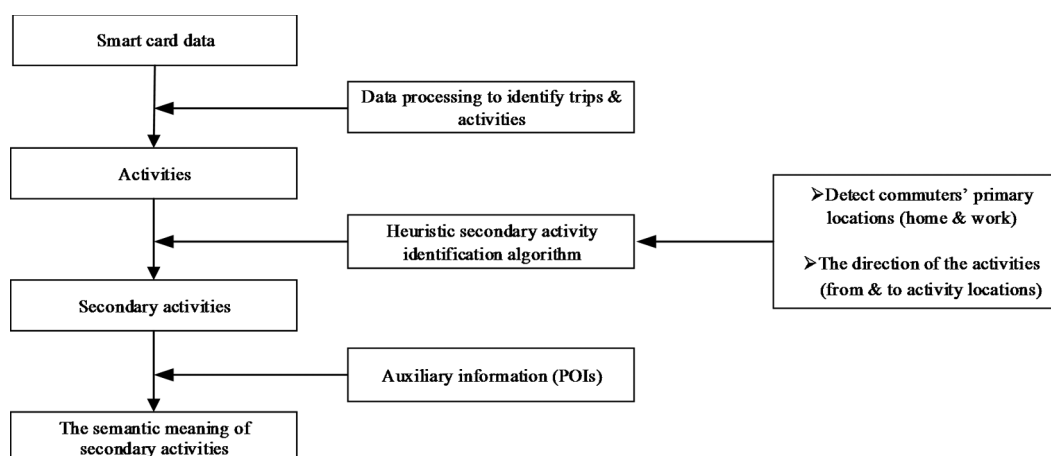


Figure 1. Logical framework.

their time within the city from smart card data and POIs respectively.

In this study, work locations have been used as an anchor to identify secondary activities. Alternatively, it is also possible to describe secondary activities using home locations, where activities would be classified as 'before-home' activities and 'after-home' activities.

2.1. Extract user activity

Before extracting activity, data is pre-processed via two exclusive steps to ensure that only effective records are selected from the large dataset:

- i) a single one-way trip from one station to another station, does not provide enough information to describe an activity. These single trips are excluded from the data (Chakirov and Erath 2012).
- ii) trips with missing key data attributes such as start time, end time, start station and end station are also excluded.

Figure 2 illustrates an activity at an individual level in terms of space and time. Spatially, (S_i) represents a station with the station number (i). T_j denotes a one-way trip from one station to another. Temporally, an activity (A) is time

(t_i) spent between two consecutive trips (T_j & T_{j+1}). Activities are further classified as either primary activities or secondary activities.

Activity identification comprises two scenarios. The first scenario examines consecutive trips where the alighting station of the first trip is the same as the boarding station of the next trip, which implies that $S(i+1)=S(i+2)$. The second scenario relaxes the station condition of the consecutive trip selection to consider the potential walking distance between two locations even if $S(i+1)\neq S(i+2)$. A pre-calculated distance matrix of walking distance between each set of stations (S_i) is used to determine whether there is an activity between two consecutive trips. Additionally, if the time spent from alighting station to boarding station is less than the expected activity time threshold (See Section 3.2.1), then the duration spent at this station is not defined as an 'activity' but a 'transfer'.

2.2. Heuristic secondary activity identification algorithm

In this study, before applying a heuristic secondary activity identification algorithm, two types of information are

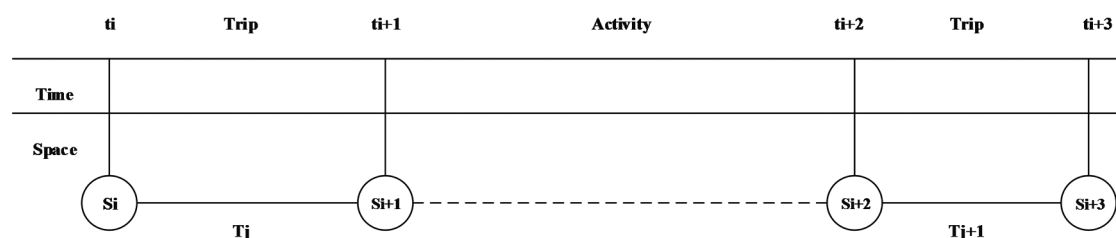


Figure 2. Schematic diagram of the data generation process for an activity. Two consecutive trips (T_j and T_{j+1}) generates an activity (A) and is calculated as $((t_{i+2})-(t_{i+1}))$ for further analysis.

required: primary locations (home & work) for each individual, and the direction (from & to) of the activity relative to those locations.

2.2.1. Detecting commuters' key locations (home & work)

Several studies have been performed on the detection of commuters from smart card data (Chakirov and Erath 2012; Zou et al. 2016). In this study, anchor points are used to create trip chains for each individual for each day (section 2.2.3.). Commuters' locations and activities provide regular patterns in the dataset revealing their travel behaviours. Here we use a similar principle to identify commuters and their home and work stations (Sari Aslam, Cheng, and Cheshire 2019; Wang et al. 2017).

Home locations are identified using the boarding stations of the first trip and the alighting stations of the last trip of an individual user on a given day. If both stations are the same or lie in a spatial proximity threshold, the stations are considered as home locations and are further analysed using the visit frequency. A similar heuristic approach is carried out to identify work locations. Consecutive trips (T_j & T_{j+1}) for all working days are evaluated. Activity location is identified using the alighting and boarding station of the first and second trip, respectively. If the selected stations match, they are further analysed for visit frequency (more than five times) and stay time duration (more than 5 hours). Nevertheless, the flexible work locations in public transport were considered using a regularity parameter (visit-frequency) for each individual, which means if the person had more than one home or work location, both locations were considered in aggregated results. More details are presented in Sari Aslam, Cheng, and Cheshire (2019). The impact of flexible work locations on secondary activities is similar. For instance, having two different work locations creates two different

after-work activities for the individual. The total count of after-work activities was used for both locations in the aggregated results.

2.2.2. From & to activity locations

In this study, to understand secondary activities from smart card data perspective, activity from & to locations are defined as the nearest spatial information relating to an activity. After defining the primary locations of each individual, the chain of activities is investigated based on the direction of travel based on (activity from-activity to) relative to those anchor locations.

From-activity location (FL) is defined as the last location before an activity. To-activity location (TL) is defined as the next location after an activity. The 'from & to locations' of an activity is translated into a binary vector based upon the location types (home, work and other).

Figure 3 illustrates 'from & to activity locations' of a secondary activity for an individual. In this example, both the locations match to WL. The activity pattern suggests a midday activity where an individual travelled from work to carry out an activity and returned to work afterwards. 'From & to activity locations' are extracted for each activity in a day for all individuals for further analysis.

2.2.3. Extracting secondary activities

To identify secondary activities in Figure 4, first, select an individual for a particular day. Second, identify all of that day's activities. Third, derive 'from & to locations' of the primary activities. As a result, secondary activities are categorized as 'before-work', 'midday', 'after-work' and 'undefined'. After completing all days for the selected individual, the next individual's data is considered using the same process.

Before-work activities are defined as taking place between the home and work locations. If an individual

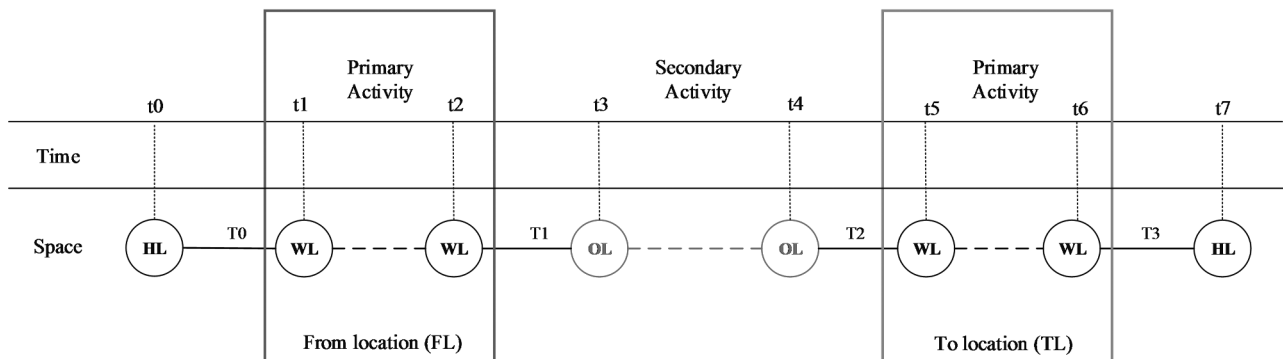


Figure 3. Schematic diagram illustrating activity chains in a day for an individual. In this case, secondary activity is marked in red at other location (OL) described using activity 'from & to locations' (FL & TL). HL and WL refer to home and work locations as primary activities, respectively.

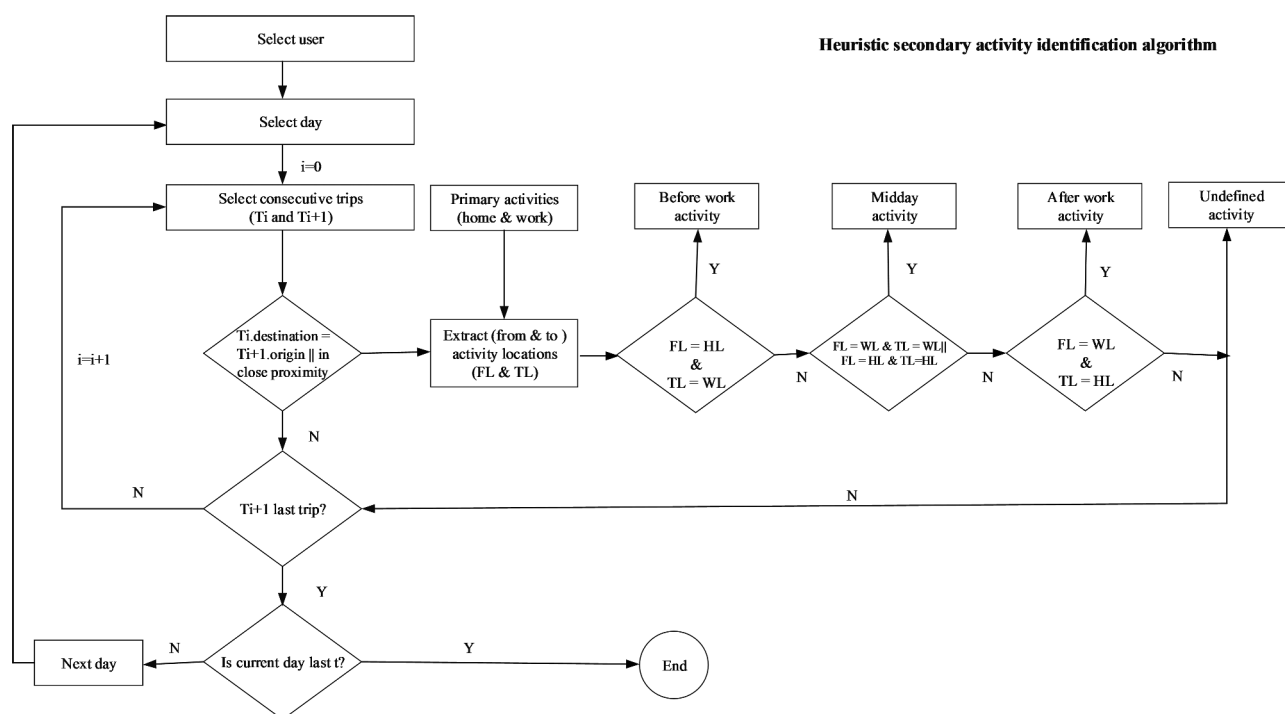


Figure 4. Flowchart of secondary activity identification algorithm for each user (FL and TL refer to from & to activity locations, respectively. HL = home locations, WL = work locations for each user).

came from home and spent time at that location before arriving at the work location, that activity is defined as 'before-work activity'. If an individual travelled for activity from the work location and returned to the work location or if he/she travelled from the home location and returned to the home location after the activity, the activity is labelled as 'midday activity'. If an individual came from the work location and spent some time at another location before going to the home, that activity is labelled as 'after-work activity'. Finally, if the activity doesn't match any defined criteria strictly, which means the nearest spatial point relating to the activity (from/to) does not have complete anchor information (defined as home or work), then the activity is labelled as 'undefined activity'. An undefined activity can have either one anchor point (ie Home – Others, Work – Others, Others – Home, Others – Work), or none (Others – Others).

In the literature, only one type of secondary activity is investigated from public transport, which is after-work activities using time constrains (Wang et al. 2017). The first constrain is calculated as the threshold of the earliest time of departure from work, and the second constrain is the finishing time of the tube lines, restricting individuals' after-work activities to before mid-night. However, in our study, secondary activities are extracted using anchor points as well as the direction information of those locations. Therefore, the proposed algorithm can capture individual-level starting and ending working

hours (flexible working hours) as a holistic picture with before-work and midday activities as captured in travel surveys (Rasouli and Timmermans 2015).

2.3. Enriched secondary activities

Location-based POIs from Foursquare or Twitter data provide ways of inferring human movements and activities (Chaniotakis, Antoniou, and Pereira 2016; Rashidi et al. 2017; Bantis and Haworth 2019; Zhu et al. 2020). To assist the inference process, large spatial-temporal transport data are commonly coupled with auxiliary information such as land use and POIs (Points of Interest), which can provide information type of performed activities (Noulas et al., 2015; Gong et al. 2016), and thus facilitate inference tasks such as activity prediction and activity pattern classification (Hasan and Ukkusuri 2014).

In activity-based modelling, primary activities such as home, work and schools are used for long-term forecasts and usually considered 'mandatory activities', the least flexible in terms of scheduling, while secondary activities are mainly considered 'maintenance activities' (dropping and picking up children and shopping) or 'discretionary activities' (eating out, entertainment, social visits, other recreational activities and doctor visits) (Castiglione, Bradley, and Gliebe 2015). We find similar activities

within the smart card dataset. However, from a land-use policy perspective, the main objective is to optimize the use of city centres to prevent congestion or areas becoming deserted at certain times. This is achieved by controlling for operating hours through planning permission (Montgomery 2017). As work trips as well as eating, shopping and entertainment trips are impacted by varying establishments' opening and closing hours, they are the most appropriate measures for demand forecast within cities (Alsger 2017). In contrast, visiting a park or bridge and social visits are less time-dependant and generate inconsistent trips on the transport network, and this discretionary character is not enough to warrant policy measures within cities (Alsger et al. 2018). Our POI dataset has similar activities, including opening and closing hours. Therefore, we have considered eating, entertainment, shopping, work, and others (travel & transport, outdoor & recreation, and home) as sub-categories of POIs for the enrichment of the secondary activities at the tube/train station level.

POIs from foursquare data are categorized based on industry classification of the visited place and easy-to-determine trip purposes (Rashidi et al. 2017). They do offer some advantages compared to other data sources such as land use. For instance, the total number of check-ins can be used to assign different weights within a trip purpose inference model. Using opening and closing hours of POIs assists in presenting urban flow within cities (Rashidi et al. 2017). In this study, we have matched 'when' attributes from smart card data with POIs to refer to secondary activities. Reducing irrelevant POIs using time-variable represent meaningful inference of the secondary activities (please see bar charts in Figure 7). The following steps have been taken:

i) *Identify the temporal characteristics of each POI*: Each POI is assigned opening and closing hours for weekdays and opening and closing hours for weekends. The most frequent opening and closing hours from Monday to Sunday are checked and assigned as the value for weekdays and weekends, respectively. If there is no information available for a particular POI, such as a Turkish restaurant, the most frequently used opening and closing hours of other Turkish restaurants are checked and assigned.

ii) *Spatial & temporal filtering*: A catchment area is defined around each station and the starting and ending hours of each secondary activity are used to filter POIs based on their opening and closing hours. For instance, an individual's before-work activity starts at 09:00 and finishes at 11:00, POIs for this activity is filtered using the opening and closing hours of POIs. If there is no overlapped information based on time attributes, those POIs such as restaurants, museums or night clubs are

excluded for further analysis. That is applied for each activity from smart card data to POIs in order to control over-representation of activity types such as eating.

iii) *Station profile using the weighted average (WAI)*: After step two, activities are grouped under their categories (eating, entertainment and shopping, work (offices and schools), and others), which is denoted as:

$$S_i = [A_{i_{eat}}, A_{i_{ent}}, A_{i_{shop}}, A_{i_{work}}, A_{i_{others}}], \quad (1)$$

The total count of activity at a specific station is defined as:

$$A_{i_{total}} = [A_{i_{eat}} + A_{i_{ent}} + A_{i_{shop}} + A_{i_{work}} + A_{i_{others}}]. \quad (2)$$

To obtain a better description of the station's characteristics, the weighted average (WAI) of each activity in station i is calculated. In Eq 3, only WAI_{eat} was presented:

$$WAI_{eat} = (A_{i_{eat}} * 100 / A_{i_{total}}) * (A_{i_{total}} / \sum_{i=0}^{n=0 \dots n} A_{i_{total}}) \quad (3)$$

Thus, the scaled activity values in Equation 1 can be replaced as:

$$S_i = [WAI_{eat}, WAI_{ent}, WAI_{shop}, WAI_{work}, WAI_{others}]. \quad (4)$$

Based on equation 4, each station has five weighted values according to its location (station)'s POIs profile.

Nevertheless, despite the wide range of positive applications, POIs from foursquare data have a number of limitations in terms of contribution bias, which means a small number of users are responsible for a substantial part of the check-ins specifically for the eating and shopping activities compare to work activities. This creates over-representation of some locations in cities (Rashidi et al. 2017). Besides, the data also suffer from demographic biases, which means the application mainly popular for younger users between 15 and 30 compared to older age groups (Longley and Adnan 2016).

3. Case study

3.1. Data and study area

London has one of the most comprehensive public transport networks in the world. Founded in 1863, London Underground, also known as Tube, is the oldest underground passenger railway network in the world covering 400 km with 270 stations.

3.1.1. Smart card data

The focus of the case study is applying the proposed model to the smart card data provided by Transport for London (TfL). The Oyster card holds the travel pass and credit for trips carried out on the TfL Oyster network. The smart card records of 10,000 TfL individuals were

selected randomly for the case study. After removing single trip users, 9900 individuals' data consisting of 1,823,906 complete trip records were considered for the examination. The unlabelled data was prepared for an individual user by extracting attributes of their daily movements such as boarding and alighting time, boarding and alighting station, and transport mode.

To evaluate the outcome of our secondary activity identification model, the ground truth smart card travel records of 40 volunteers were gathered, along with the information about their home and work locations: 8,156 trip records (approximately 4,000 data points) covering two months. In the labelled dataset, we have the demographic details of the users in addition to the information available through the smart card automated fare collection system: anonymous identifier of the users, journey timestamp, boarding and alighting stations and type of activities. This includes the classification of before-work, midday, after-work and undefined activities as well as sub-classification of eating, entertainment, shopping, work, and other activities.

3.1.2. Point of interest (POIs)

Points of interest data for this study was collected using the Foursquare Location API. The total number of POIs from Foursquare users captured in London around tube/train stations (walking distance 800 m) is about 38,921,981, and the total number of check-ins is 81,328,352.

The POIs include a broad classification of location category and various types, as shown in Table 1. Additional attributes in the dataset are working hours, working days. The percentage of each activity type in the dataset is eating (24%), entertainment (18%), shopping (17%), travel & transport (16%), outdoor & recreation (12%), work (12%) and home (1%).

Table 1. Activity types from foursquare data for the study.

Activity Category	Activity Location Type
Work	Schools, government buildings, offices, post office, colleges and universities, social club, TV station, warehouse and other places
Eating	Coffee shop, sandwich, pizza, cafe, diner, bakery, burger, restaurant, steakhouse, breakfast, bagel shop, etc.
Entertainment	Art, pub, nightclub, theatre, entertainment, club, bar, concert hall, other nightlife, opera house, casino, event space, dance studio etc.
Shopping	Supermarket, store, pharmacy, mall, boutique, plaza, miscellaneous shop, farmers market, automotive shop, food & drink shop, bookstore etc.
Others	Outdoor & recreation (park, playground, bridges, ski areas etc.) and travel & transport (roads, bus stops, tube stations, bike rental/bike share points, airports etc.)

3.2. Results

3.2.1. Extracting activities between consecutive trips

In this study, two scenarios are taken into consideration spatially when identifying the activity between two trips: where the distance is 0 metres ($S_i + 1 = S_i + 2$), or where the distance is greater than 0 metres but less than 800 metres ($S_i + 1 \neq S_i + 2$). Based on the first scenario, a total of 234,371 unique activities were identified. Whereas, on relaxing the consecutive trip condition to allow up to 800 metres walking distance (RTPI 2018), a total of 249,863 activities were identified. The second scenario represented an improvement of 6% in the identification of secondary activities. In literature, recently, walking distance (800 m) and transfer time threshold (60 mins) were used to investigate trip chaining assumptions (Alsger et al. 2018). As the last step, activity identification was investigated using the transfer time threshold. The average transfer time of some of London stations was estimated by TfL (2019) as approximately 20 min. In this study, a transfer time threshold of 15 min was applied, 4962 records were identified as transfers activities and were excluded from the activity dataset.

3.2.2. The temporal characteristics of secondary activities

The analysis of smart card data was employed to find the variation in secondary activities of commuters. Figure 5 illustrates the comparison of secondary activities (almost 30% in unlabelled smart card data, and 28.1% in labelled smart card data) in both datasets. While unlabelled smart card data capture more activity, labelled data (section 3.2.2.) provides detail on the nature of the secondary activities. The highest and lowest activity counts in both datasets are after-work and before-work activities, respectively. Undefined activities account for 7.23% in unlabelled data and almost 5% in labelled data, where we can observe more information about the nature of these activities. For instance, 1.8% and 0.94% are labelled as social visits (the activity locations are far from the centre of the city such as home locations) and holidays (the activity locations are airports) correspondingly. The rest of the undefined activities are labelled as shopping (0.81%), entertainment (0.44%), eating (0.38%), work (0.31%) and other activities (0.25%) such as walking in the city or park, and doctor or other appointments.

Although the algorithms are defined without using the time variable as a characteristic, identified activities still have a temporal characteristic such as boarding-alighting time and duration of the activity, as well as the day of the

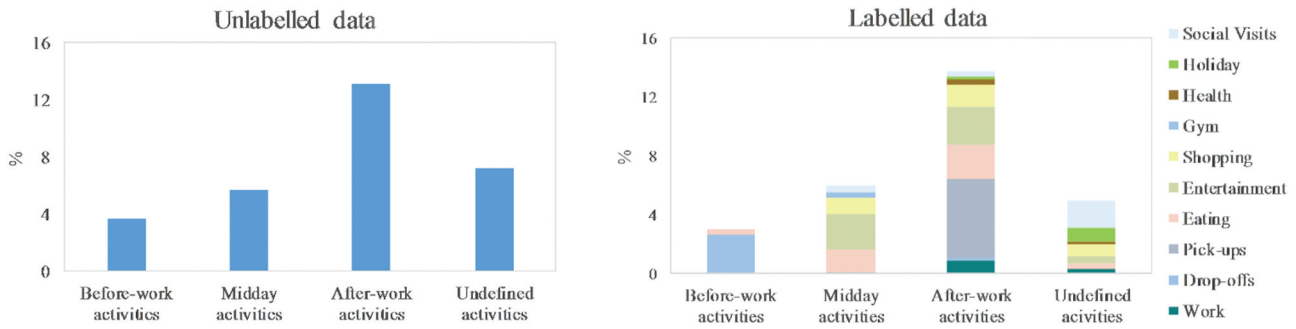


Figure 5. Unlabelled and labelled data using the classification of secondary activities.

activity. Therefore, the temporal characteristics of secondary activities are investigated further in details.

Figure 6 illustrates an aggregate analysis of secondary activities and their characteristics, such as activity duration in each day (heat maps), and activity start and end time (line charts).

The first column presents before-work activities. The heat map of before-work activities highlights a consistent two to three hours window during the weekdays. These activities are less significant during the weekends. In addition, the start and end times peak between 08:00 to 09:00 (blue line), and 10:00 to 12:00 (red line) respectively.

The second column represents midday activities. The heat map illustrates a consistent window of two to four hours during the weekdays and two to five hours' during the weekends. There are two peaks in the total count of start and end hours. The first and second peaks of the start hours are 12:00 and 16:00, respectively. The first and second peaks of the end hours are 14:00 to 18:00. The smaller peaks, appearing almost three hours later, might be due to home-to-home midday activities, especially during the weekends

The third column presents after-work activities. The heat map of after-work activities shows that there is some difference in activity duration between weekdays and weekends. After-work activities are confined to a two-to-four-hour window during the weekdays. However, after-work activities appear throughout a longer period during the weekend and don't present as consistently as weekdays. This can be seen in the most intense colour on Saturdays compared to Sundays. Also, the line charts show a different pattern of start and end hours compared to before-work activities. The count of start hours of the after-work activities reaches a peak from 15:00 to 17:00. However, the counts of end hours show two peaks around 19:00 and 23:00. The starting of after-work activities may be regular due to fixed departure times from work, especially during the weekdays. Ending of after-work activities is irregular due to the different time taken to reach home locations.

The last column shows the temporal variation of undefined activities. Duration from the heat map is less than five hours during the weekdays compared to more than five hours for weekends, especially on Saturdays. The start hours of undefined activities present three

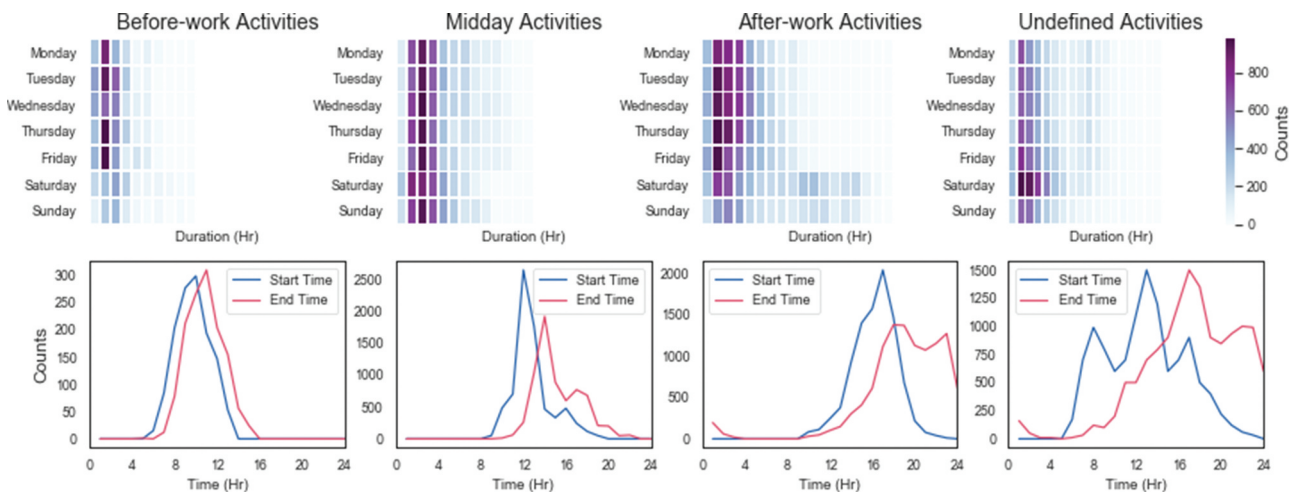


Figure 6. The secondary activities (before-work, midday, after-work and undefined) are presented during the whole week.

peaks which are 08:00, 12:00 and 16:00. The end hours of undefined activities present three peaks which are 11:00, 17:00 and 23:00. The reason for these three peaks is that they share an anchor point from one of the key locations such as home or work in the dataset (3.66%).

As a result, duration is an important characteristic in defining activities (Chakirov and Erath 2012; Zou et al. 2016), and the duration of primary activities, in the literature, is defined as from ten to fifteen hours and six to nine hours for home and work activities respectively (Chakirov and Erath 2012; Devillaine, Munizaga, and Trépanier 2012; Zou et al. 2016; Sari Aslam, Cheng, and Cheshire 2019). To the best of our knowledge, this study is the first to define the duration of the secondary activities, which are four hours or less, especially during the weekdays. On the contrary, some studies directly used time constraints to extract secondary activities (Wang et al. 2017). However, individuals' start and end hours of secondary activities present temporal variation. Thus, in this study, the sequence of activity chains for each individual is used to have an accurate estimate for travel purposes, and the result of the analysis are presented at an aggregate level.

3.2.3. Validation of the identified Secondary Activities

Comprehensive validation of the activities identified from the smart card data is difficult to achieve due to the limited availability of the test or survey data. Two validation approaches have been used to see the accuracy of the proposed algorithm.

The first approach is to gather the accuracy from the ground truth. Two months of the trip and trip-purpose data (as mentioned in section 3.1.1.) with approximately 4,000 data points are used for the validation purposes to see the result of the proposed algorithm. 80% accuracy for after-work, 76% accuracy for before-work, almost 70% accuracy for midday and 57% accuracy for undefined activities were obtained using the proposed secondary activity identification algorithm.

The second method is to use another model as a baseline (as mentioned in section 2.2.3.) with which to compare the accuracy of after-work activities only (Wang et al. 2017). The estimation of after-work activities using the baseline approach is only accurate by 67.5% due to the earliest departure time from work being set as 16:00. However, almost 20% of the after-work activities are labelled in the dataset as children pick-ups from school during 15:00–17:00. Child pick-up/drop off duties are mentioned by the related literature (Castiglione, Bradley, and Gliebe 2015; Xiao, Juan, and Zhang 2016). Besides, the baseline can be extended to include before-work and midday with time constraints

of 07:00 to 09:00 and 12:00 to 14:00, respectively. This will yield a success rate of only 62% and 56% respectively using the same validation dataset. Thus, the proposed algorithm provides better identification of secondary activities and demonstrate a complete picture compared to the existing baseline model, which help meaningful enrichment for a dynamic city e.g. London.

3.2.4. The semantic meaning of secondary activities

The number of check-ins around stations under each of the five categories (eating, entertainment, shopping, work and others) are investigated for the 626 train/tube stations across London, using smart card data and POIs (see section 2.3).

Figure 7 illustrates the aggregated analysis of an individual trip's purpose according to where and why people spent their time within the city. Each secondary activity is explained using the three charts, reading anti-clockwise: First, the peak locations of secondary activity are presented in the London map using only smart card data. Second, the percentages of activity types from the total counts of POI check-ins are illustrated for each secondary activities in the bar chart. Finally, both information such as identified secondary activities and their enrichment form POIs are presented for the selected central London stations such as Oxford Circus, Piccadilly Circus.

First, the high count of before-work activity stations is illustrated. As well as central London stations, some residential and school stations are highlighted such as Richmond, Clapham Common and Hampstead. From the count of check-ins from POIs, it is inferred that work activities (work and school) are the main type of before-work activity, which is about 42% as the highest probability. The basis for this inference is twofold: first, work activity locations (section 2.2.1.) are identified using duration times more than 5 hours (Sari Aslam, Cheng, and Cheshire 2019). Therefore, some part-time workers' work activity which is less than 6 hours, might be captured as before-work activity. Second, student activities (pick-ups and drop-offs) are also highlighted as work activities in this study since TfL does not have student card information for children under 11 years old. (see section 1). Therefore, we expected that most of the drop off activities which activity-based models have mentioned in early studies (Castiglione, Bradley, and Gliebe 2015) would appear here under work activities. The main activity at the majority of the selected central London stations (except for Green Park and Bond Street) are also inferred to be work activities.

The high count of midday activity in stations mainly in central London and Stratford are taken to represent office workers' lunch breaks. Due to home-to-home

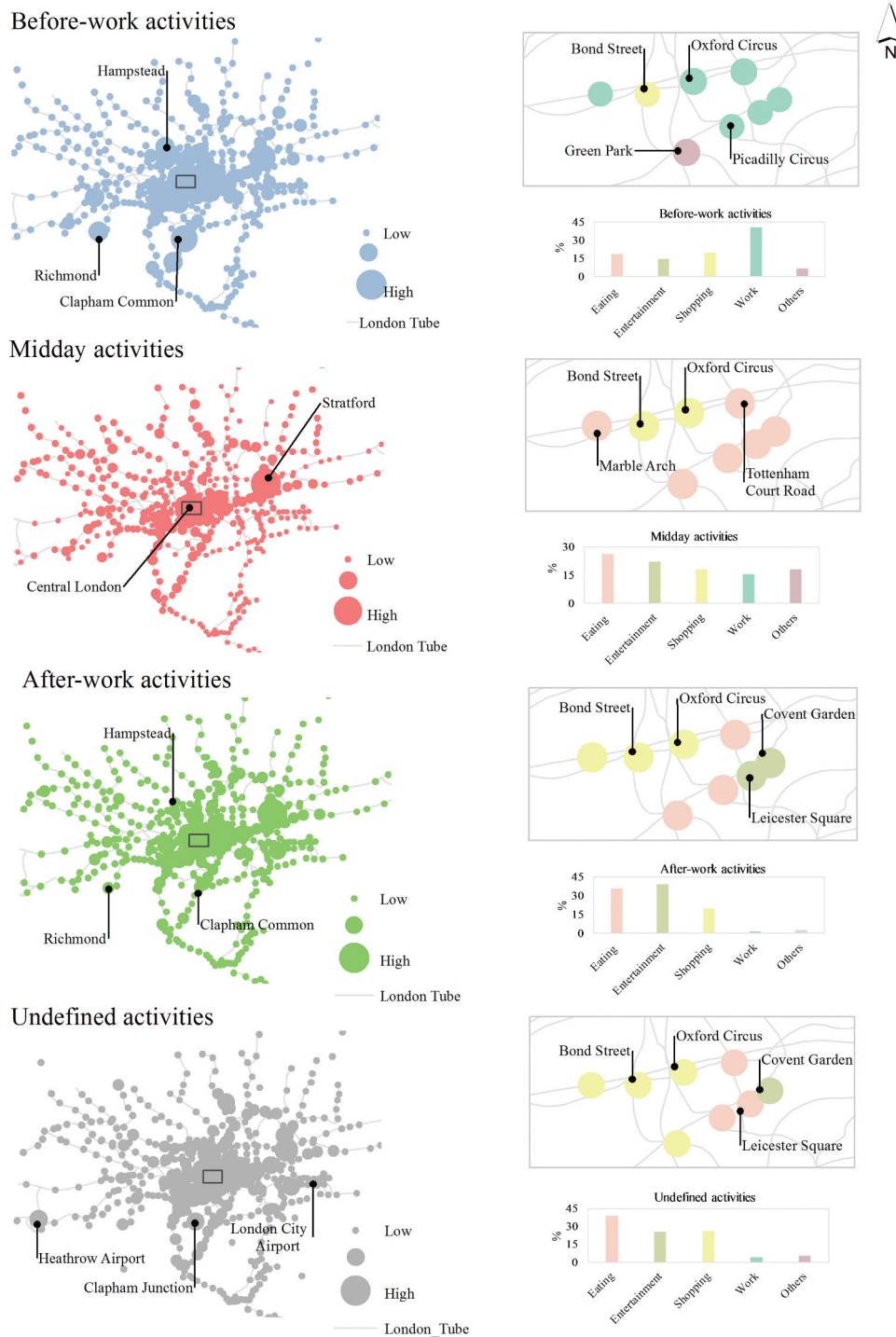


Figure 7. Secondary activities at the station level with POIs are presented to infer the semantic meaning of secondary activities. Bar charts illustrate the total count of check-ins from all London stations.

middy activities, especially during the weekends, the bar chart illustrates not only eating activities but also other activities such as shopping and entertainment. In addition, once we looked at the central London stations, the purpose of the majority of midday activities was referred to as 'eating' except for Oxford Circus and Bond Street, which were stated as 'shopping'.

The high counts of after-work activity stations on the map suggest that there is an overlap with both before-work and midday activities; this combination can also be seen from labelled data (section 3.1.1.). Although the total count of after-work activities in residential and school locations is similar to that for before-work activities, the total number of after-work identified activities

overall (13.14%) is more than the total number of both before-work (3.72%) and midday (5.71) activities combined (section 3.2.2.). This suggests that the biggest contribution comes from entertainment (40%), eating (34%) and shopping (almost 19%) activities rather than activities at school locations, which can be seen from the bar chart as well as the central London stations. Furthermore, the selected London stations show that the close stations have similar inferences under a certain category. For instance, the inferred activity at Covent Garden and Leicester Square is entertainment while the inferred of activity at Oxford Circus and Bond Street is shopping.

Finally, the high counts of undefined activity locations show that almost 2% of the undefined activity appears either at interchange stations or London airports. A few examples are highlighted in the London map. The first reason for this 15 min transfer time might be less for those transport hubs in metropole cities. Besides, some studies excluded those interchange stations as a step before defining primary locations (Li et al., 2015). However, most of the interchange stations in London have large spaces for passengers to spend their time while they are waiting for their journey. Hence, the study provides the station-based enrichment as well, even though the total count is presented simply as eating, entertainment or shopping in the bar chart. Finally, the selected central London station activity is classified as eating and shopping. The reason might be those undefined activities have a spatial point from one of the anchor points (mentioned in 3.2.3.) such as work locations. For instance, the individual comes from work may use a different mode of transport to go back home such as car or bike.

The secondary identification algorithm is able to highlight before-work, midday or after-work activities as locations strictly. Most of the activity detection and inference models are assigned using the highest probability of POIs as an attractiveness proxy (Gong et al. 2016; Wang et al. 2017; Alsgar et al. 2018). However, in this study, the starting and ending hours of secondary activities are compared with the opening and closing hours of POIs for each location before assigning the highest probability of land-use POIs. Hence, Figure 7 has provided meaningful enrichment.

Furthermore, the same London stations are enriched by different activity types during the day using the classification of secondary activities. For instance, Leicester Square is inferred as a work location under before-work activities, eating location under midday and entertainment location under after-work while Marble Arch is inferred as a work location under before-work activities, eating location under midday activities

and shopping location under after-work activities. That shows how incorporating secondary activities with dynamic POIs (opening and closing hours and user check-ins) may lead to a meaningful activity inference.

As a result, the framework uses big data sources to investigate individual secondary activities to refer to travel purpose. The approach demonstrates how secondary activities can be derived from smart card data using the proposed secondary activity identification algorithm. The spatio-temporal characteristics of secondary activities for each individual have quantified in the aggregate analysis that the majority of the secondary activities are four hours or less, especially during the weekdays. The study presents how secondary activities are combined with POIs using spatial and temporal filtering, which help the meaningful inference of secondary activities even though POIs have a number of limitations such as contribution bias and demographic bias (see section 2.3.). As a result, the purpose of travel is different for the same stations and individuals during different times of day in dynamic cities, which assist in representing urban flow in a more accurate and complete picture. The outcome is beneficial for urban and infrastructure/transport planners to develop sustainable cities.

4. Conclusion

The large volume of individual-level smart card data present opportunities to generate new insights into travel behaviour research and urban modelling. This study aimed to demonstrate a framework specifically for enriching the semantics of secondary activities by combining smart card data with additional Points of Interest (POIs) in a complex urban environment. A heuristic model was proposed to derived travel activities from smart card data, define primary home and work locations, and identify secondary activity based on from & to locations of activities, with ancillary POI data to estimate the likely nature of the secondary activity.

First, the proposed secondary activity identification algorithm can detect meaningful locations for secondary activity types with high precisions. A 'heuristic secondary activity identification algorithm' was applied to tube/train travellers in London and the algorithm reaches accuracies of 80% for after-work activities, 76% for before-work activities, and 70% for midday activities based on volunteers' responses. Thus, the high-level classification of the activities helps better understanding of travel behaviour of users and facilitates efficient and sustainable urban transport systems development.

Secondly, a framework integrating the secondary activity identification algorithm with auxiliary socio-

functional information was introduced to investigate the reason for the travel in the case of regular users. In the case of London, the identified secondary activities were enriched using Foursquare data. Five semantic categories (work, eating, entertainment, shopping and others) of POIs with opening and closing hours find different spatial patterns for the various activity types. Hence, linking human travel behaviour with urban functions demonstrate how trip purposes and urban mobility patterns can be beneficial for city planners.

Lastly, the proposed method provides a meaningful way to understand individuals' activities from a data-driven perspective. As an alternative to the traditional travel demand survey, this work offers a cost-effective approach for human mobility. Future work will build on this foundation to explore secondary activities further including undefined activities. We believe that semi-supervised learning methods have the potential to advance mobility analysis in large cities using a combination of limited labelled data from volunteer surveys and unlabelled data from smart cards. Additionally, bus journeys, representing a major exclusion from this study, can be included if missing alighting information is inferred as part of the activity identification process.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the Economic and Social Research Council [1477365] as part of the Consumer Data Research Centre (CDRC) project (ES/L011840/1). Besides, thanks to Transport for London (TfL) and Dr Juntao Lai for sharing London's Oyster card data and Foursquare POIs check-ins data respectively.

ORCID

Nilufer Sari Aslam  <http://orcid.org/0000-0001-8858-5514>
 Di Zhu  <http://orcid.org/0000-0002-3237-6032>
 Tao Cheng  <http://orcid.org/0000-0002-5503-9813>
 Mohamed R. Ibrahim  <http://orcid.org/0000-0001-7733-7777>
 Yang Zhang  <http://orcid.org/0000-0003-1524-385X>

References

Alsger, A. A., A. Tavassolia, M. Mesbah, L. Ferreira, M. Hickman. 2018. "Public Transport Origin-Destination Estimation Using Smart Card Fare Data." *Sante publique (Vandoeuvre-les-*

- Nancy, France)* 28 (3): 391–397. doi:10.1017/CBO9781107415324.004.
- Alsger, A. A. M. 2017. "Estimation of Transit Origin-destination Matrices Using Smart Card Fare Data." MSc thesis, The University of Queensland.
- Arentze, T., and H. Timmermans. 2007. "Robust Approach to Modeling Choice of Locations in Daily Activity Sequences." *Transportation Research Record: Journal of the Transportation Research Board* 2003 (1): 59–63. doi:10.3141/2003-08.
- 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.
- Bantis, T., and J. Haworth. 2019. "Non-Employment Activity Type Imputation from Points of Interest and Mobility Data at an Individual Level: How Accurate Can We Get?" *ISPRS International Journal of Geo-Information* 8 (12): 560. doi:10.3390/ijgi8120560.
- Ben-akiva, M., J. L. Bowman, and D. Gopinath. 1996. "Travel Demand Model System for the Information Era." *Transportation* 23: 241–266.
- Castiglione, J., M. Bradley, and J. Gliebe. 2015. *Activity-Based Travel Demand Models: A Primer*. Washington, DC: Transport Research Board. (accessed June 18, 2020).
- Chakirov, A., and A. Erath. 2012. "Activity Identification and Primary Location Modelling Based on Smart Card Payment Data for Public Transport." *13th International Conference on Travel Behaviour Research*, (July).
- Chaniotakis, E., C. Antoniou, and F. Pereira. 2016. "Mapping Social Media for Transportation Studies." *IEEE Intelligent Systems* 31 (6): 64–70. IEEE. doi:10.1109/MIS.2016.98.
- Chu, K. K. A., and R. Chapleau. 2010. "Augmenting Transit Trip Characterization and Travel Behavior Comprehension." *Transportation Research Record: Journal of the Transportation Research Board* 2183 (1): 29–40. doi:10.3141/2183-04.
- 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 (1): 48–55. doi:10.3141/2276-06.
- Farooqi, H., M. Mesbah, and J. Kim. 2018. "Applications of Transit Smart Cards beyond A Fare Collection Tool: A Literature Review." *Advances in Transportation Studies* 45 (July): 107–122. doi:10.4399/978255166098.
- Gong, L., X. Liu, L. Wu, Y. Liu. 2016. "Inferring Trip Purposes and Uncovering Travel Patterns from Taxi Trajectory Data." *Cartography and Geographic Information Science* 43 (2): 103–114. Taylor & Francis. doi:10.1080/15230406.2015.1014424.
- Gordon, J., H. N. Koutsopoulos, N. H. M. Wilson, J. P. Attanucci. 2013. "Automated Inference of Linked Transit Journeys in London Using Fare-transaction and Vehicle Location Data." *Transportation Research Record: Journal of the Transportation Research Board* 2343 (1): 17–24. DOI:10.3141/2343-03.
- Goulet-Langlois, G. 2016. *Exploring Regularity and Structure in Travel Behavior Using Smartcard Data*. Massachusetts Institute of Technology. <https://pdfs.semanticscholar.org/dd24/51f990de5ce5325c73975e149c26f04c77f0.pdf>
- Hasan, S. and S. V. Ukkusuri. 2014. 'Urban activity pattern classification using topic models from online geo-location data'. *Transportation Research Part C Emerging*

- Technologies, 44:363–381. Elsevier Ltd. doi:doi: 10.1016/j.trc.2014.04.003
- Kuhlman, W. 2015. *The Construction of Purpose Specific OD Matrices Using Public Transport Smart Card Data*. Delft University of Technology. <https://repository.tudelft.nl/islandora/object/uuid%3A7190712e-0913-4849-89ae-d1a1a88e66d2>
- Li, G., L. Yu, W. Wu, W. S. Ng and S. T. Goh. 2015. "Predicting Home and Work Locations Using Public Transport Smart Card Data by Spectral Analysis." *IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC 2015-October*: 2788–2793. doi:10.1109/ITSC.2015.445.
- Longley, P. A., and M. Adnan. 2016. "Geo-temporal Twitter Demographics Geo-temporal Twitter Demographics." *International Journal of Geographical Information Science* 30 (2): 369–389. Taylor & Francis. doi:10.1080/13658816.2015.1089441.
- Lu, X., and X. Gu. 2011. "The Fifth Travel Survey of Residents in Shanghai and Characteristics Analysis." *Urban Transport of China* 9: 1–7.
- Ma, X., Y.-J. Wu, Y. Wang, F. Chen, J. Liu. 2013. "Mining Smart Card Data for Transit Riders' Travel Patterns." *Transportation Research Part C: Emerging Technologies* 36: 1–12. Elsevier Ltd. doi:10.1016/j.trc.2013.07.010.
- Montgomery, J. 2017. *The New Wealth of Cities: City Dynamics and the Fifth Wave*. London and New York: Taylor & Francis. <https://g.co/kgs/wNk2Rj>
- Nakamura, K., F. Gu, V. Vichiensan, H. Yoshitsugu. 2016. "Failure of Transit-Oriented Development in Bangkok from a Quality of Life Perspective." *Asian Transport Studies* 4 (1): 194–209. doi:10.1590/S0100-72032012000400008.
- Noulas, A., B. Shaw, R. Lambiotte, C. Mascolo. 2015. "Topological Properties and Temporal Dynamics of Place Networks in Urban Environments." *arXiv:1502.07979 [Physics]*: 431–441. doi:10.1145/2740908.2745402.
- 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. Elsevier Ltd. doi:10.1016/j.trc.2010.12.003.
- Pinjari, A. R. N. Eluru, S. Srinivasan, J. Y. Guo, , R.B. Copperman, I. N. Sener, C.R. Bhat. 2007. "CEMDAP: Modeling and Microsimulation Frameworks, Software Development, and Verification CEMDAP." (June 2014).
- Rashidi, T. H., A. Abbasi, M. Maghrebi, S. Hasan, T. S. Waller. 2017. "Exploring the Capacity of Social Media Data for Modelling Travel Behaviour: Opportunities and challenges." *Transportation Research Part C: Emerging Technologies* 75: 197–211. Elsevier Ltd. doi:10.1016/j.trc.2016.12.008.
- Rasouli, S., and H. Timmermans. 2015. "Activity-based Models of Travel Demand : Promises, Progress and Prospects." 5934 (December). *International Journal of Urban Sciences*. doi:10.1080/12265934.2013.835118.
- RTPi. 2018. *How Far Is It Acceptable to Walk ?* https://www.rtpi.org.uk/media/2739252/wyg_gareth_pdf.pdf
- Saif, M. A., M. M. Zefreh, and A. Torok. 2019. "Public Transport Accessibility : A Literature Review." *Periodica Polytechnica Transportation Engineering Public* 47 (1): 36–43. doi:10.3311/PPtr.12072.
- Sari Aslam, N., and T. Cheng. 2018. "Smart Card Data and Human Mobility." In *Consumer Data Research*, edited by P. Longley, J. Cheshire, and A. Singleton, 111–119. London, UK: UCL Press. <https://www.jstor.org/stable/j.ctvqhsn6.11>
- Sari Aslam, N., T. Cheng, and J. Cheshire. 2019. "A High-precision Heuristic Model to Detect Home and Work Locations from Smart Card Data." *Geo-spatial Information Science* 22 (1): 1–11. Taylor & Francis. doi:10.1080/10095020.2018.1545884.
- TfL. 2019. *Out-of-station Interchanges*. Accessed 26 September 2019. <https://tfl.gov.uk/corporate/publications-and-reports/out-of-station-interchanges>
- Wang, Y., G. H. D. A. Correia, E. de Romph, 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 (January): 40–47. Elsevier. doi:10.1016/j.jtrangeo.2017.06.010.
- Xiao, G., Z. Juan, and C. Zhang. 2016. "Detecting Trip Purposes from Smartphone-based Travel Surveys with Artificial Neural Networks and Particle Swarm Optimization." *Transportation Research Part C: Emerging Technologies* 71: 447–463. Elsevier Ltd doi:10.1016/j.trc.2016.08.008.
- Yang, Y., A. Heppenstall, A. Turner, A. Comber. 2019. "Who, Where, Why and When? Using Smart Card and Social Media Data to Understand Urban Mobility." *ISPRS International Journal of Geo-Information* 8 (6): 271. DOI:10.3390/ijgi8060271.
- Zhang, Y., and T. Cheng. 2020. "A Deep Learning Approach to Infer Employment Status of Passengers by Using Smart Card Data." *IEEE Transactions on Intelligent Transportation Systems* 21 (2): 617–629. IEEE. doi:10.1109/TITS.2019.2896460.
- Zhang, Y., T. Cheng, and N. Sari Aslam. 2019. "Exploring the Relationship between Travel Pattern and Social - Demographics Using Smart Card Data and Household Survey." In *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 10–14, Multidisciplinary Digital Publishing Institute (MDPI). The Netherlands: Enschede.
- Zhong, C., S. M. Arisona, X. Huang, M. Batty, G. Schmitt. 2014. "Detecting the Dynamics of Urban Structure through Spatial Network Analysis." *International Journal of Geographical Information Science* 28 (11): 2178–2199. doi:10.1080/13658816.2014.914521.
- Zhu, D., F. Zhang, S. Wang, Y. Wang, X. Cheng, Z. Huang, Y. Liu, et al. 2020. "Understanding Place Characteristics in Geographic Contexts through Graph Convolutional Neural Networks Understanding Place Characteristics in Geographic Contexts through Graph Convolutional Neural Networks." *Annals of the American Association of Geographers* 110 (2): 408–420. Routledge. doi:10.1080/24694452.2019.1694403.
- Zhu, D., N. Wang, L. Wu, Y. Liu. 2017. "Street as A Big Geo-data Assembly and Analysis Unit in Urban Studies: A Case Study Using Beijing Taxi Data." *Applied Geography* 86: 152–164. Elsevier Ltd. doi:10.1016/j.apgeog.2017.07.001.
- Zou, Q., X. Yao, P. Zhao, H. Wei, H. Ren. 2016. "Detecting Home Location and Trip Purposes for Cardholders by Mining Smart Card Transaction Data in Beijing Subway." *Transportation* 3. Springer US. doi:10.1007/s11116-016-9756-9.