Routledge
Taylor & Francis Group

Transportation Planning and Technology

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/gtpt20

Alighting location estimation from public transit data: a case study of Shenzhen

Nilufer Sari Aslam, Joana Barros, Han Lin, Roberto Murcio & Honghan Bei

To cite this article: Nilufer Sari Aslam, Joana Barros, Han Lin, Roberto Murcio & Honghan Bei (24 Jul 2024): Alighting location estimation from public transit data: a case study of Shenzhen, Transportation Planning and Technology, DOI: 10.1080/03081060.2024.2382247

To link to this article: https://doi.org/10.1080/03081060.2024.2382247

Published online: 24 Jul 2024.

TRANSPORTATION PLANNING AND TECHNOLOGY
https://doi.org/10.1080/03081060.2024.2382247

Alighting location estimation from public transit data: a case study of Shenzhen

Nilufer Sari Aslam[a,b], Joana Barros[a,b], Han Lin[c], Roberto Murcio[a,b] and Honghan Bei[c,d,e]

[a]Department of Geography, Birkbeck, University of London, London, UK; [b]Centre for Advanced Spatial Analysis, University College London, London, UK; [c]Collaborative Innovation Centre for Transport Studies, Dalian Maritime University, Dalian, People's Republic of China; [d]School of Maritime Economics and Management, Dalian Maritime University, Dalian, People's Republic of China; [e]School of Management, Shanghai University, Shanghai, People's Republic of China

ABSTRACT

This study proposes a framework to estimate alighting locations from Smart Card Data (SCD) that are absent information on entry-only public transport systems such as buses and trams. The proposed method uses the characteristics of SCD to (i) determine boarding locations from SCD and GPS-bus data based on exact match and time windows using common attributes, (ii) infer individuals' home locations and user types from multimodal SCD, (iii) estimate alighting locations using inferred information with different scenarios such as with and without home locations based on the type of users. Reliable results are obtained once home locations are identified with high confidence for all user types. The proposed framework is applied to Shenzhen, China as a case study to validate the proposed model's effectiveness. The study offers valuable insight into aligning location estimation from user types to optimise the quality of public transport planning and services.

ARTICLE HISTORY
Received 8 October 2023
Accepted 12 July 2024

KEYWORDS
Alighting location estimation; inferring user types; home location identification; trip chaining; smart card data

1. Introduction

Automated fare collection systems (AFCs) offer a valuable source of information for planning transport services and facilities in urban environments (Anda, Erath, and Fourie 2017). AFCs collect users` travel information through smart cards (SC) in two systems (Hussain, Bhaskar, and Chung 2021). The first is entry-exit systems, which collect the tap-in and out information from passengers at boarding and alighting stops. In this system, the travel fare depends on the distance, such as in Queensland, Australia (Alsger et al., 2016b) and Seoul, Korea (Lee et al. 2021). The second is entry-only systems, where users must only tap in once when entering the transport system. Such systems are generally preferred in flat-fare systems, where the fare is independent of

CONTACT Nilufer Sari Aslam n.aslam.11@ucl.ac.uk Department of Geography, Birkbeck, University of London, London WC1E 7HX, UK; Centre for Advanced Spatial Analysis, University College London, London W1T 4TJ, UK

the journey's distance, for instance, in New York (Barry, Freimer, and Slavin 2009) and Chicago (Zhao, Rahbee, and Wilson 2007) in the USA, or Guangzhou (Yu and He 2017) and Shenzhen (Yan, Yang, and Ukkusuri 2019) in China. While entry-exit systems produce accurate spatiotemporal information on boarding and alighting locations, entry-only systems do not record alighting locations, which restricts distinguishing travel time, transfer time and time spent between two consecutive journeys. That causes severe problems in trip chaining/travel behaviour methodology to understand current and future demand, including service facilities in public transport networks (Anda, Erath, and Fourie 2017; Hussain, Bhaskar, and Chung 2021).

Heuristic and Machine Learning are two main approaches for estimating passengers' alighting locations from public transport networks. The heuristic (rule-based) approach uses a set of rules/algorithms known as the trip chaining model to identify alighting locations from SCD. The model relies on two main assumptions: (i) The destination of the current trip is the same as the origin of the next trip if the passenger has more than a single trip per day (Barry et al. 2002), and (ii) The passenger's last trip's destination is the same as the origin of the first trip on a relevant day (Barry et al. 2002) or the origin of the first trip on the next day (Munizagaa and Palma 2012). The heuristic method later on incorporates the idea of walking distance in the trip-chaining methodology (Zhao, Rahbee, and Wilson 2007; Hofmann and Mahony 2005; Trépanier, Tranchant, and Chapleau 2007; Wang, Attanucci, and Wilson 2011; Alsger et al. 2018) and transfer time (Zhao et al.; Nassir et al. 2011; Ma et al. 2013; Alsger 2016a; Sari Aslam et al. 2020) to improve alighting location estimation (Cerqueira, Arsenio, and Henriques 2023), except using home locations from smart card data. Even though the rule-based approach is widely used in literature due to the unavailability of survey data, the Machine Learning (ML) approach is also applied without relying on predetermined equations or rules. Jung and Sohn (2017a) estimated the destination of bus journeys combining entry-only SCD and land use characteristics using a deep learning model for Seoul, Korea. Yan, Yang, and Ukkusuri (2019) proposed a two-step algorithm to determine boarding stops and estimate alighting stops for bus systems in Shenzhen, China. Assemi et al. (2020) estimated alighting stops inference accuracy using the characteristics of SCD with Artificial Neural Networks (ANN). Although the ML approach allows the estimation of alighting locations with high accuracy, the method requires model training from travel surveys, which are unavailable for many cities – particularly those in the Global South (Ordóñez Medina 2018; Yang et al. 2019).

Recent studies highlight the importance of user types for travel behaviour and activity patterns (Goulet-Langlois, Koutsopoulos, and Zhao 2016). Identifying user types with alighting location estimation from SCD benefits planning of transport services delivered by public agencies, such as planning bus locations and the optimisation of public transit timetables to a wide range of user types (He, Trépanier, and Agard 2021). While user types, as a card type, sometimes exist from SCD (Munizaga and Palma 2012; Tao, Rohde, and Corcoran 2014; Chen and Fan 2018), some studies need to extract this information to analyse alighting location estimation. A particular focus has been on analysing mobility patterns distinguishing regular (Hasan et al. 2013) and irregular users (Yang et al. 2019) as the basis for estimating travellers' upcoming movements from SCD. Despite a body of literature on the travel behaviour of other user types (Ghaemi et al. 2017; He, Trépanier, and Agard 2021; Hussain, Bhaskar, and Chung 2021), those with

disabilities, students (Cong, Gao, and Juan 2019) and older adults/special users, to the best of our knowledge, has not widely been included into approaches for alighting estimation from SCD.

This study proposes a framework to estimate alighting locations using the heuristic approach with additional information extracted from SCD, i.e. individuals' home locations and user types such as adults (below 60 years old), students (between the ages of 6 and 14 and full-time students), and older adults/special users (above 60 years old adults and users with valid certificates to have fully discounted journeys such as disable users, etc.) Determining boarding locations by combining SCD-bus with bus-GPS data, the framework estimates alighting locations from multimodal SCD (subway and bus).

The contributions of the study are summarised as follows:

- The proposed methodology provides an enhancement for alighting location estimation using extracted information from smart card data, such as individuals' home locations and user types as an alternative to the conventional travel surveys.
- The alighting location algorithm demonstrates how users' home location estimations with high confidence are beneficial in estimating individuals' last locations in travel behaviour research.
- An empirical study using Shenzhen SCD (subway and bus) and GPS-bus data validates the proposed model's effectiveness while highlighting the current challenges and limitations in determining boarding and alighting locations in the study area.
- The last location estimation with user types (adults, students and older adults/special users) has the potential to open new revenues to plan bus stop locations and the amenities around the bus stop locations for public transport planning.

The next section of the study presents the research framework and methods proposed for this study. Then, the results are presented as a case study in Shenzhen, China, in Section 3. The challenges and limitations of the study are discussed in Section 4. Finally, Section 5 concludes the findings and future directions of the study.

2. Data and method

2.1. Study area and data

Shenzhen is a fast-growing region in China known for its financial services and high-tech industries. The study area covers 2000km^{2}and houses a population of about 17 million today.

Shenzhen's transport system in 2014, when the dataset used for this study was collected, was five lines and 118 stations with a volume exceeding 3.94 million passengers on its busiest day (Yang et al. 2019). Bus and subway trips accounted for 89% of total passenger trips, a considerable proportion of daily trips in the study area. Besides, Shenzhen had a flat-fare system, and the fare was independent of the journey's distance.

Two data types are available for the study: Smart Card, which contains data on subway/train and bus journeys, and GPS dataset on buses. These are detailed below.

The Shenzhen SCD covers the period from 9 June 2014–13 June 2014 (Monday to Friday). The SCD contains 36,786,796 journey records with 4,172,621 individuals

(anonymised). The modes of public transport are under two categories, such as subway and bus, consisting of 29,37% and 54,97% of the total journeys in the period, respectively. In addition, the dataset contains combined trains and bus journeys, constituting more than 15% of the total journeys. The attributes of SCD-bus include user ID, time stamps, bus plate number, bus route number, direction (up-down), the start time of the trips/journeys, the amount of money available and the amount of money left after travel in each card.

The GPS-bus data was used as an additional source to investigate bus journeys in conjunction with SCD. The data, which covers the same period (9 June 2014–13 June 2014), consists of individual bus trajectories and includes bus plate number, bus line (route) number, the coordinates of bus locations, and time stamp. A total of 151,631,764 GPS records are included in this dataset.

2.2. Method

The proposed methodological framework consists of six steps from multimodal transport data illustrated in Figure 1. The method starts with data cleaning to improve the data quality and accuracy of the estimated values from SCD while reducing the noise from unprocessed SCD (Dacheng et al. 2018). It removes duplicates and incomplete records (such as missing information regarding boarding/alighting time or stations, missing user IDs, etc.) from both datasets, i.e. SCD and GPS-bus. In addition, single trips (per day) are excluded from the SCD since they need to provide more information for individuals to estimate destination information in public transport systems. (Alsger 2016a; Gordon et al. 2013; Hora et al. 2017; Jung and Sohn 2017a; Kumar, Khani, and He 2018; Ma et al. 2013; Munizagaa and Palma 2012; Munizaga et al. 2014; Nunes, Dias, and Cunha 2015; Trépanier and Chapleau 2006). Note that the same individuals' rest of the trips are still used for the following analysis.

The second step focuses on determining boarding locations from SCD with the help of GPS-bus data required for only-entry systems when boarding location (Farzin 2008; Gordon et al. 2013; Wang, Attanucci, and Wilson 2011) or boarding location and time (Lahat, Adali, and Jutten 2015) are unavailable. In addition, often, there are no reliable timetables for bus lines, as it is the case for Shenzhen (Yan, Yang, and Ukkusuri 2019). Therefore, there is a need to infer the required information by fusing SCD with other data sources, such as bus line data (Barry, Freimer, and Slavin 2009; Liu, Tan, and Liu 2020) or GPS-bus data (Chen and Wang 2013; Huang et al. 2020).

In this study, SCD-bus boarding stations are determined using bus-GPS data as follows: (i) common attributes from SCD and GPS-bus data (date, bus plate number, route number, and start time of the trips with bus-GPS records) are matched. In practice, the exact match using common attributes between bus-SCD and bus-GPS data cannot always be captured due to mismatches in the time stamps on the two datasets or missing records from bus-GPS data. (ii) The start time of the bus journey from SCD is further used with the initiating of time windows. Time windows, such as + windows (adding minutes) and – windows (subtracting minutes) from bus-SCD, were adopted, incrementally increasing and decreasing values of the start time of bus journeys for a higher match rate between the two datasets. For instance, the start time of the bus journey in the SCD is 8:00 am, yet the time values from GPS-bus data are 7:58 am

Figure 1. The workflow of the framework.

and 8:06 am. The time windows are created from + minutes and – minutes of the start time of bus journeys, and 7:58 am is used from bus GPS data to identify the closest point to the boarding location from SCD. This is because GPS records can have inaccuracies ranging from 10 to 100 metres (Huang et al. 2020), hence using + and – windows allows for improving match rates displayed in Figure 2.

After determining boarding bus locations using bus-GPS data, user types were extracted from the SCD in step 3 using the payment policy given by Shenzhen Transport (Shenzhen Metro Group Co 2022). In this step, three types of information on fare arrangements, such as the amount of money in each card per individual (T_{money}), the amount of money remained in the card after travel (T_{deal}) and the mode of transport on public transport, are used for classifying user types for alighting location estimation from SCD. After calculating how much each user paid for a trip (T_{Paid}) based on the mode of transport, the discount rate (D) is calculated as the proportion of paid trips (T_{Paid}) to the full cost of the same trips (T_{Cost}). The details of the payment policy for

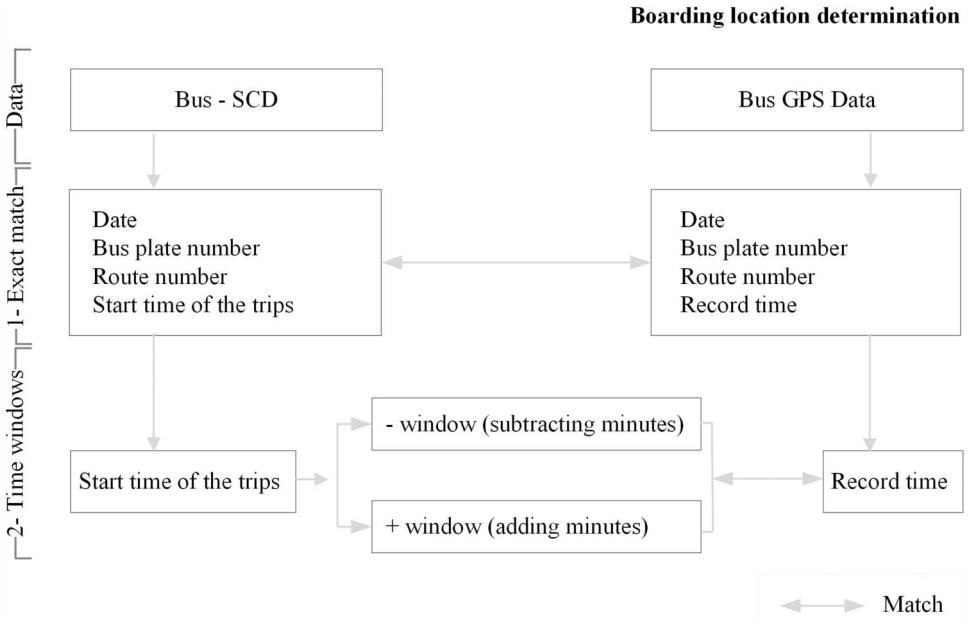

Figure 2. The flowchart of boarding location determination using time windows.

the case study are given in Section 3.1.

$$T_{Paid} = (T_{money} - T_{deal}) \tag{1}$$

$$D = \left(\frac{T_{Paid}}{T_{Cost}}\right) * 100 \tag{2}$$

The estimation of the alighting stations is based on the assumptions from the literature: the passengers' last trip destination is the same as the origin of the first trip on a relevant day from the entry-exit systems (Barry et al. 2002) or the origin of the first trip on the next day from only-entry systems (Munizagaa and Palma 2012). This assumption was complemented with walking distance parameters to estimate the last locations (Zou et al. 2016; Alsger et al. 2018; Yan, Yang, and Ukkusuri 2019). Based on those assumptions, in step 4 of this study, the home station locations per individual are extracted. It's important to note that this is a limitation set by the SCD, which only permits home locations to be represented at the station level. The first journey's origin is picked with the first journey's origin on the next day. If both stations are the same, selected stations are further analysed using frequency parameters. If both stations differ, nearby stations are determined using walking distance first (less or equal to 1000 m – (following Yan, Yang, and Ukkusuri 2019), then the stations that are deemed frequent (i.e. reached the frequency parameter) are inferred as a home location for that user (following Sari Aslam, Cheng, and Cheshire 2019). In this case, even if passengers' last location information is unavailable, knowing the home station location determines the last location information in nearby areas with frequency parameters. The home will be referred to as home location in the remaining of the paper.

The next step focuses on alighting location estimation, illustrated in Figure 3. After selecting individuals, the bus journey for the first day is selected as a current journey and checked against the next bus journey. If the next bus journey is available, it assumes that the current journey's end location is the same as the next journey's start location (Barry et al. 2002). If the next journey is unavailable, the current journey is assumed to be the last journey for the day. The last journey is dealt with in three stages in Figure 3. These are Stage 1 – If home location is available, the home location is assigned as the individual's last-day location. This is important because inferred individuals' home locations are added to the richness of the input data. Stage 2 – If the user's last location for the day is still unavailable, the bus route number for the last and first journeys is checked. If the bus route numbers match and the journeys are in opposite directions (up-down), the current journey's end station is assigned as the first journey's start location (Nassir et al. 2011). By including the information on bus route numbers and directions, this stage further fills the data gaps that were not addressed in Stage 1. Stage 3 – Otherwise, the current journey's end location is assigned to the first journey's start location in Stage 3 (Barry et al. 2002). Once the destination of the last journey is defined, the algorithm checks the next bus journey during the day and other days for the same individuals and the rest of the users in the dataset.

Although alighting locations are estimated from the individuals' daily journey patterns once the start and end locations are the same or are near each other (Trépanier and Chapleau 2006; Wang, Attanucci, and Wilson 2011; Gordon 2012; Alsger et al., 2016b; Nunes, Dias, and Cunha 2015; Yan, Yang, and Ukkusuri 2019), an additional assumption for the last location estimation, i.e. Stage 1, which applies to individuals for which the home locations are identified using additional frequency parameter, is compared to the assumptions without home locations, i.e. Stage 2 and Stage 3 in this study.

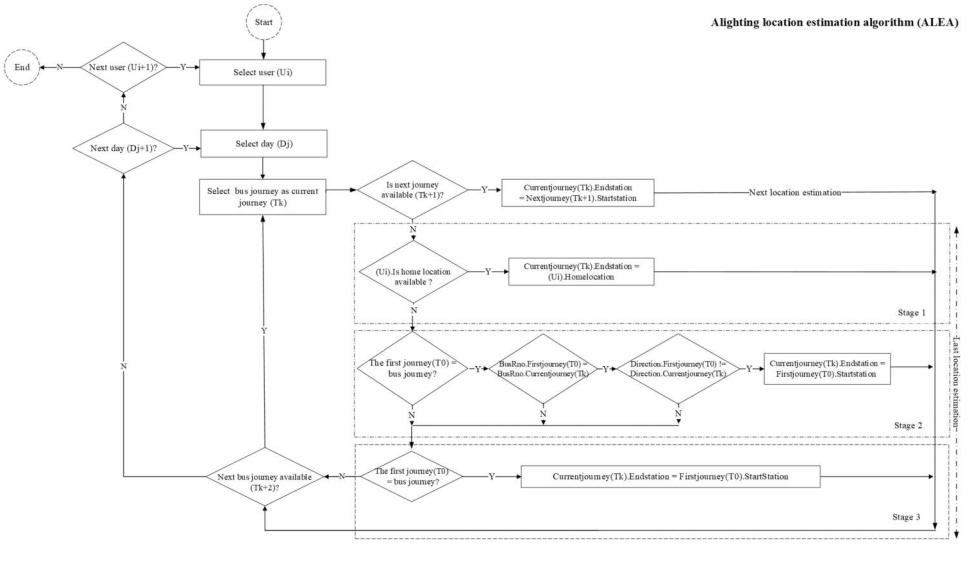

Figure 3. The flowchart of the alighting location estimation algorithm (ALEA). The last location estimation is illustrated under Stage 1 (with home locations), Stage 2 and Stage 3 (without home locations).

The last step evaluates the assumptions using available destination locations from subway journeys due to the unavailability of ground truth data. The accuracy of the algorithm applied to the data in Stage 1 with home locations is compared to the accuracy of the algorithm based on data that does not include the home location in Stage 3 (baseline assumption proposed by Barry et al. (2002)). Besides, different frequency parameters are employed to provide the required confidence level with/without home locations.

3. Results

Data cleaning, such as removing duplicates and missing values, is summarised in Figure 4 for both SC and GPS-bus datasets. While journey counts decreased during the removal of duplicates, the number of unique users remained unchanged in SCD. After excluding missing information, nearly 77% and 89% of data are left in the SCD and bus-GPS data, respectively. The data cleaning process excludes single trips that is 2,464,142 of journey records (8.76%) and 809,258 of the individuals (19.39%) from SCD. Note that, the remaining trips for the same individuals were kept and used for the analysis. At the end of the data cleaning process, 25,677,544 journeys (69% of the journeys), 3,353,116 individuals (80% of the individuals' data) from SCD and 136,154,881 records (89% of the total) from bus-GPS data were available for the analysis.

To estimate alighting time and locations for entry-only systems, determining bus boarding locations is an essential step for trip chaining methodology. 15,406,526 bus-SCD from 25,677,544 SCD journey counts do not have boarding locations but contain time stamps in the dataset. Two types of data matches are applied between bus-SCD

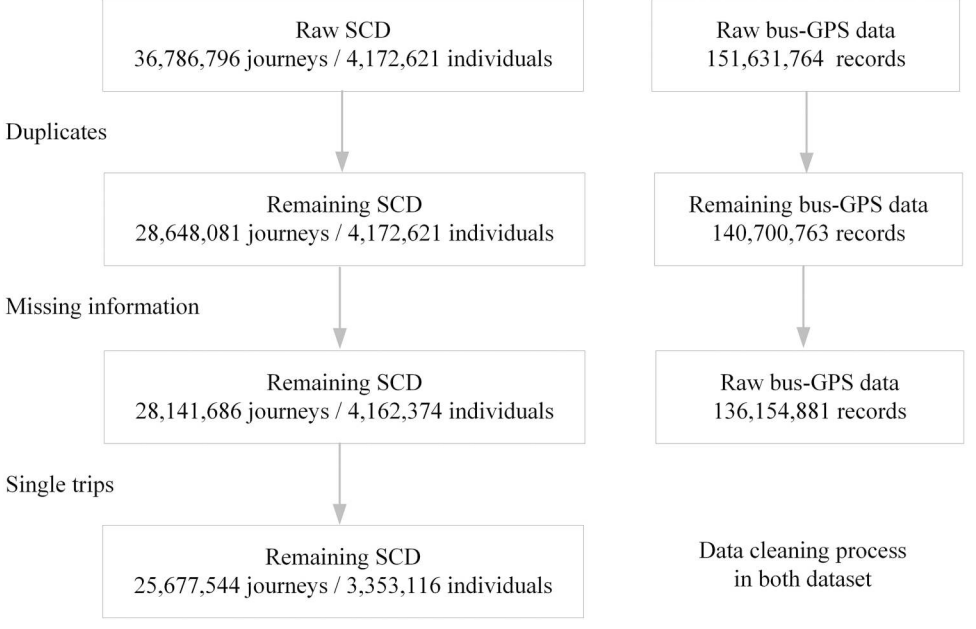

Figure 4. Data cleaning process with raw and remaining datasets.

and bus-GPS data. 59% of the boarding locations of bus journeys, i.e. 9,089,850 records, are captured by matching the common attributes from both datasets, such as date, bus plate number and route. The remaining records from bus-SCD, based on the start time of the journeys, are further examined within a window that incorporates both ` + ` and `-` minutes, such as 2 minutes (creating a total window of 4 minutes), 3 minutes (total of 6 minutes), and 4 minutes (total of 8 minutes). Employing only '+' windows of 2 , 3 , and 4 minutes captures the match rate to 62%, 70%, and 71%, respectively. Utilising both '+' and '-' windows for the same range slightly improves the match rate, achieving 67%, 71%, and 72%, respectively. These variations in match rates indicate that GPS data records do not refresh every 20–40 s, and the absence of GPS data points accounts for the inaccuracies in boarding estimation within the system. The factors for the missing GPS data may vary, such as poor signal, device or data collection errors, etc. The data matched in the 3-minute and 4-minute windows is similar, with the 4-minute window showing a slight advantage. Consequently, the 4-minute window was selected for the analysis, which is 11,092,699 records at the end of the data match process for boarding location estimation.

3.1. Extracted information from SCD

One of the extracted information from SCD is user types, identified with three main attributes, such as the available amount of money for the travel per individual (T_{money}), the amount of money left in the card after travel (T_{deal}) for each journey and mode of transport (subway, bus, or combined subway and bus). After calculating how much each user paid for a trip (T_{Paid}), the discount rate (D) is used for identifying user types from the passengers' discount policy given by Shenzhen Municipal Bureau of Transportation (Shenzhen Metro Group Co 2022). According to the policy, full-pay users (below 60 years old), referred here as 'adults', have a 5% and 20% discount for subway and bus journeys, respectively. The rate for children (between the ages of 6 and 14 and older full-time students), here referred to as 'students', is 50% on bus and subway journeys. Lastly, older adults/special discount users, which include adults aged 60 and above, users with valid certificates due to retirement, disability, and Shenzhen Municipal Riding Card holders, who are eligible for fully discounted fares as key workers (i.e. fire-rescue personnel), have free passes via the subway and buses. Note that a transfer time within 90 minutes gives passengers 0.4% off, a discount that applies to students and adults (Shenzhen Metro Group Co 2022).

The chart in Figure 5 shows the distribution of user types based on modes of transport extracted from SCD using the payment discounts. Adults represent 55% of bus, 22% of train and 17% of (subway and bus) journeys. Students use the bus (4%) and subway (1%) at much lower levels. Older adults/special users makeup only 1% of the journey counts. The limitation of using fares for the user classification concerns multimodal trips (subway + bus) for students and older adults/special users. Due to the significant discount applied before transfer trips, the fare system does not record the trip's second (or further) legs. As a result, the classification does not capture the subway and bus journeys for those user types.

The second extracted information from SCD is the home locations for each user. Home locations are extracted for commuters using attributes such as the first location,

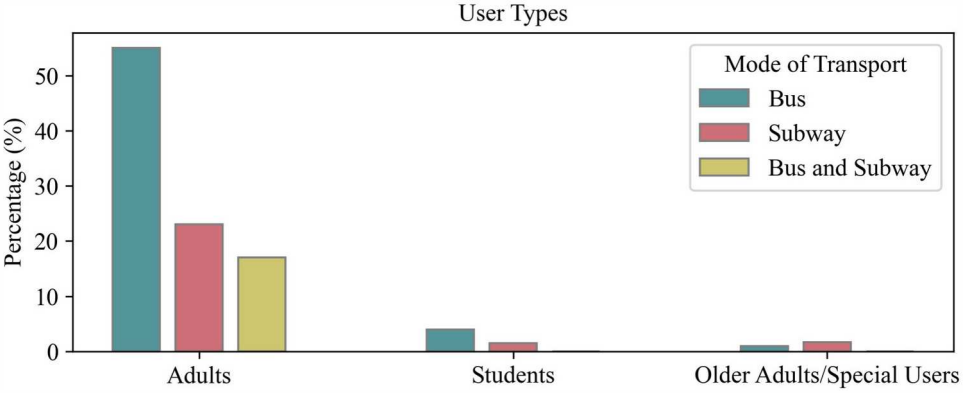

Figure 5. Identified user types based on the mode of transport using journey counts.

walking distance (less or equal to 1000 m – following Yan, Yang, and Ukkusuri 2019), and trip frequency. The frequency parameter is the main difference from alighting location assumptions based on the question, 'How many times has the user visited this location?'

According to existing research, identifying home travel patterns often involves using a high-frequency threshold, such as four and five times on working days (Hasan et al. 2013; Huang et al. 2018). Conversely, low-frequency thresholds are typically associated with identifying travel for non-standard work schedules such as entertainment, shopping, etc. from SCD (Sari Aslam 2022).

Figure 6 presents various frequency thresholds captured for determining home locations from SCD. Decreasing the frequency (a loose condition) leads to identify more individuals, whereas increasing the threshold (a tight condition) results in fewer records in the dataset. In Figure 6, a noticeable decline occurs at frequency 3, which stabilises for frequencies four and five. This pattern suggests that almost 69% and 70% of individuals presented similar behaviour with frequencies four and five, respectively, throughout the weekdays (Monday to Friday). The selection of frequency five is based on the minimal discrepancy in user counts. Frequency counts provide reliable information for estimating alighting locations when the final destinations are unavailable from systems that only record entries

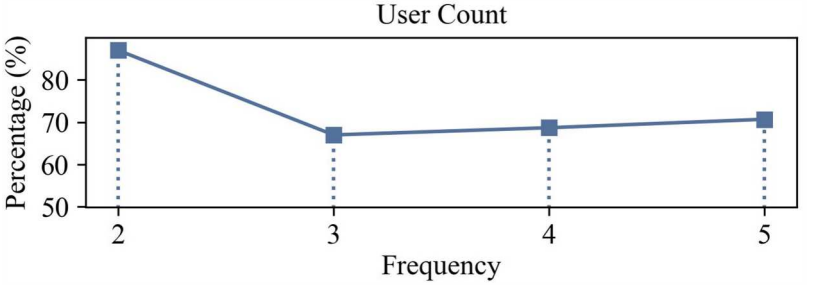

Figure 6. The frequency counts to attain confidence in identifying home locations.

3.2. The result of the various scenarios

The proposed assumptions are applied to the journey data in this section with multiple scenarios as follows: 1) identified number of alighting locations in Stage 1, Stage 2 and 3, 2) each user type, such as adults, students, and older adults/special users with home locations (Stage 1) and without home locations (Stage 2 and Stage 3) and 3) validation results with various frequency values from subway-SCD.

From 3,778,050 bus-SCD records without last location stops are examined in the trip chains. 2,520,196 records are captured using individuals' home locations with frequency 5, including 543,709 and 714,145 of the records identified from Stages 2 and 3, respectively. Since Stage 3 serves as the default condition, ensuring the last location matches the first in daily journey pairs, there are no unidentified records in the dataset.

These findings are further investigated based on user types and illustrated in Figure 7A. According to the results, alighting locations are estimated for adults at 52.48% from Stage 1, 3.27% from Stage 2 and 38.67% for Stage 3. Besides, the detection rate for students is captured at 2.76% from Stage 1, 0.65% from Stage 2 and 1.18% from Stage 3. The detection rate for older adults/special users is 0.99% for Stage 1 and 0.5% for Stage 3. There is no detection rate for older adults/special users from Stage 2 due to the complete discounted journeys. This result shows that the alighting location identification with home locations (Stage 1) provides a higher detection rate than Stages 2 and 3 for all user groups.

The validation of the study is followed on subway SCD due to the lack of ground truth data and the results are shown in Figure 7B. The same assumptions, i.e. Stage 1 and Stage 3, are applied to subway SCD evaluating the last location estimation. The results are validated from the destination points of subway journeys. Note that Stage 2 requires bus route number and direction information and is unavailable from subway SCD. As a result, 90.05% for frequency three, 91.42% for frequency four, and 92.15% for frequency five are captured using home locations in Stage 1. In comparison, 81.01% accuracy was obtained without home locations in Stage 3. The validation results show that more confidence in the accuracy of home locations provides a better outcome for the last location estimation from SCD. Second, start and end stations are not always the same, even for

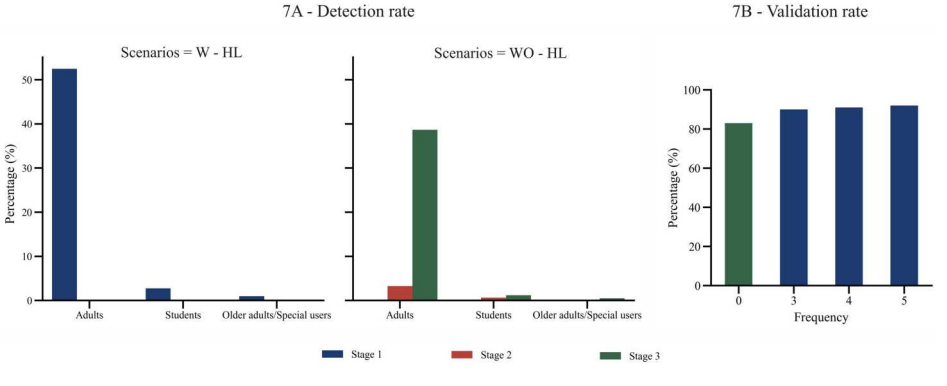

Figure 7. Detection rate (A) from bus journey data and validation rate (B) from subway journey data are identified in each Stage (W-HL and W0-HL refer to alighting location estimation with and without home locations, respectively).

regular users. Finding the last location with the help of walking distance and frequency as the home location provides better accuracy in subway SCD.

4. Discussion

This study aimed to estimate alighting locations for entry-only systems using SCD with different scenarios, such as with/without home locations. The study highlights the importance of frequency parameters for trip-chaining assumptions with home locations using big data sources such as SCD and bus-GPS data. Both data sources provide opportunities to improve our understanding of human mobility, trip purposes and demand patterns in public transport networks, albeit with some limitations. For instance, historical data do not represent the latest status of transport networks in the study area, even though the findings are still valuable to observing the improvement of accuracy with high confidence where information about an individual's home is identified. Besides, almost 30% of the bus-SCD cannot be matched with bus-GPS data to determine bus boarding locations, and the absence of GPS data is the reason for the system's failure to estimate boarding locations.

The study proposed methodology using a heuristic approach, which uses pre-defined rules from the available dataset to reduce the use of traditional travel surveys. The proposed methodology has limitations in finding unseen patterns/scenarios to help transport planning agencies. The ML approach can handle such scenarios to estimate alighting locations accurately once travel survey / labelled data are available. However, training the ML models from conventional travel surveys may not represent the whole population in urban environments due to small sample sizes and low update frequencies, which are unavailable for the study's methodology. The improvement in data collection methodologies through smart card data, including user feedback with apps or online platforms, can facilitate the transition from heuristic approaches to ML methodologies in public transportation systems.

This study also presents the information based on user types such as adults, students, and older adults/special users. Combining user information with the last locations from bus journeys can enhance the planning of bus stops' facilities in urban environments. Understanding the distribution of user types and their last locations based on different bus stops and routes, including specific times, can provide insights into the demand patterns for specific user groups. Transport or urban planners can allocate resources efficiently to meet varying daily-peak demands. Second, knowing the distribution of different user types and their last locations from bus journeys can help to identify potential safety concerns. For instance, if there is a higher concentration of older adults/special users using specific bus stops, extra safety measures can be implemented to assist them during boarding and alighting. Last, looking at alighting location estimation from user's types can be valuable for policymakers and city planners to make informed decisions about public transportation growth and funding allocation. However, identifying user types such as students and older adults/special users may involve bias once they have significant or fully discounted journeys, e.g. older adults/special users' multimodal journeys (subway and bus).

5. Conclusion

Big data sources, such as SC and GPS-bus data, generate valuable insights into travel demand forecasting and transport planning. This study aimed to demonstrate a

framework to estimate alighting locations for only-entry systems, such as buses and trams. The proposed heuristic algorithm compares scenarios with home locations (Stage 1) and without home locations (Stage 2 and Stage 3), including user types, i.e. adults, students, and older adults/special users, without relying on travel surveys. An application for the Shenzhen case study demonstrated the proposed framework's effectiveness with the detection and validation rates obtained from individuals' bus and subway journeys. A higher detection rate for the last location estimation is obtained with home locations for all user groups. Besides, validation accuracy with home location is achieved with different confidence levels using frequency parameters. 90.05% for frequency three, 91.42% for frequency four, and 92.15% for frequency five are captured using home locations compared to 81.01% accuracy without home locations. The validation results demonstrate that more confidence in the accuracy of home locations provides a better outcome for the last location estimation from SCD.

The proposed framework can be further improved in a number of various ways. Single trips are currently excluded from the dataset. During the trip chaining, assumptions can be included and investigated regarding people's travel behaviour within public transport systems. Second, home locations with different frequencies by incorporating additional data sources, such as demographic data and land use data, can be further investigated for the distribution of different population groups such as overnight workers, etc. Last, user types and their spatial and temporal travel behaviours from bus destinations can be further investigated with additional land use data sources such as Points of Interest (POIs) to optimise bus service and facilities in cities.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the Sustainable Mobility and Equality in mega-ciTy Regions (SIMETRI) project: patterns, mechanisms, and governance, funded by ESRC Economic and Social Research Council [grant no: ES/T000287/1].

References

Alsger, A. 2016a. *Estimation of Transit Origin Destination Matrices Using Smart Card Fare Data*. The University of Queensland.

Alsger, Azalden, Behrang Assemi, Mahmoud Mesbah, and Luis Ferreira. 2016b. "Validating and Improving Public Transport Origin–Destination Estimation Algorithm Using Smart Card Fare Data." *Transportation Research Part C: Emerging Technologies* 68: 490–506. http://dx.doi.org/10.1016/j.trc.2016.05.004.

Alsger, Azalden, Ahmad Tavassoli, Mahmoud Mesbah, Luis Ferreira, and Mark Hickman. 2018. "Public Transport Trip Purpose Inference Using Smart Card Fare Data [Online]." *Transportation Research Part C*. Elsevier 87 (3): 123–137. https://doi.org/10.1016/j.trc.2017.12.016.

Anda, C., A. Erath, and P. J. Fourie. 2017. "Transport Modelling in the Age of Big Data [Online]." *International Journal of Urban Sciences*. Taylor & Francis, https://doi.org/10.1080/12265934.2017.1281150.

Assemi, B., A. Alsger, M. Moghaddam, M. Hickman, and M. Mesbah. 2020. "Improving Alighting Stop Inference Accuracy in the Trip Chaining Method Using Neural Networks." *Public Transport* 12 (1): 89–121.

Barry, J. J., R. Freimer, and H. Slavin. 2009. "Use of Entry-Only Automatic Fare Collection Data to Estimate Linked Transit Trips in New York City." *Transportation Research Record* 2112 (1): 53–61. https://doi.org/10.3141/2112-07.

Barry, James J., Robert Newhouser, Adam Rahbee, and Shermeen Sayeda. 2002. "Origin and Destination Estimation in New York City with Automated Fare System Data [Online]." *Transportation Research Record: Journal of the Transportation Research Board* 1817 (1): 183–187. http://trrjournalonline.trb.org/doi/10.31411817-24.

Cerqueira, Sofia, Elisabete Arsenio, and Rui Henriques. 2023. "Is There Any Best Practice Principles to Estimate Bus Alighting Passengers from Incomplete Smart Card Transactions?" *Transportation Research Procedia* 72: 3395–3402.

Chen, Zhen, and Wei Fan. 2018. "Extracting Bus Transit Boarding and Alighting Information Using Smart Card Transaction Data." *Journal of Public Transportation* 22 (1): 40–56. http://dx.doi.org/10.5038/2375-0901.

Chen, J., and Z. Wang. 2013. "Algorithm of Estimating Alighting bus Stops of Smart Card Passengers Based on Trip-Chain." *Applied Mechanics and Materials* 253–255: 1918–1921.

Cong, J., L. Gao, and Z. Juan. 2019. "Improved Algorithms for Trip-Chain Estimation Using Massive Student Behaviour Data from Urban Transit Systems." *IET Intelligent Transport Systems* 13 (3): 435–442. https://doi.org/10.1049/iet-its.2018.5183.

Dacheng, C., Y. Ruizhi, S. Lei, T. Kiat, D. Hui, J. K. H. Whye, and N. Kiong. 2018. "Traveler Segmentation Using Smart Card Data with Deep Learning on Noisy Labels." Proceedings of ACM KDD conference (Vol. 10).

Farzin, J. M. 2008. "Constructing an Automated Bus Origin – Destination Matrix Using Farecard and Global Positioning System Data in São Paulo, Brazil [Online]." *Transportation Research Record* 2072 (1): 30–37. https://journals.sagepub.com/doi/10.31412072-04.

Ghaemi, M. S., B. Agard, M. Trépanier, and V. Partovi Nia. 2017. "A Visual Segmentation Method for Temporal Smart Card Data." *Transportmetrica A: Transport Science* 13 (5): 381–404. https://doi.org/10.1080/23249935.2016.1273273.

Gordon, J. B. 2012. *Intermodal Passenger Flows on London's Public Transport Network: Automated Inference of Full Passenger Journeys Using Fare-Transaction and Vehicle-Location Data*. Berkeley: University of California.

Gordon, Jason B., Harilaos N. Koutsopoulos, Nigel H. M. Wilson, and John P. Attanucci. 2013. "Automated Inference of Linked Transit Journeys in London Using Fare-Transaction and Vehicle Location Data." *Transportation Research Board* 2343 (1): 17–24. https://doi.org/10.3141/2343-03.

Goulet-Langlois, G., H. N. Koutsopoulos, and J. Zhao. 2016. "Inferring Patterns in the Multi-Week Activity Sequences of Public Transport Users." *Transportation Research Part C*. Elsevier Ltd 64:1–16. https://doi.org/10.1016/j.trc.2015.12.012.

Hasan, Samiul, Christian M. Schneider, Satish V. Ukkusuri, and Marta C. González. 2013. "Spatiotemporal Patterns of Urban Human Mobility." *Journal of Statistical Physics* 151 (1-2): 304–318. http://dx.doi.org/10.1007/s10955-012-0645-0.

He, L., M. Trépanier, and B. Agard. 2021. "Space – Time Classification of Public Transit Smart Card Users' Activity Locations from Smart Card Data." *Public Transport* 0123456789.

Hofmann, M., and M. O. Mahony. 2005. "Transfer Journey Identification and Analyses from Electronic Fare Collection Data." Proceedings of the IEEE, Vienna, Austria.

Hora, Joana, Teresa Galvão Dias, Ana Camanho, and Thiago Sobral. 2017. "Estimation of Origin-Destination Matrices Under Automatic Fare Collection: The Case Study of Porto Transportation System [Online]." *Transportation Research Procedia*. Elsevier B.V 27:664–671. https://doi.org/10.1016/j.trpro.2017.12.103.

Huang, J., D. Levinson, J. Wang, J. Zhou, and Z.-J. Wang. 2018. "Tracking Job and Housing Dynamics with Smartcard Data." *Proceedings of the National Academy of Sciences* 115 (50): 12710–12715. http://dx.doi.org/10.1073/pnas.1815928115.

Huang, D., J. Yu, S. Shen, Z. Li, L. Zhao, and C. Gong. 2020. "A Method for Bus OD Matrix Estimation Using Multisource Data." *Journal of Advanced Transportation* 2020.

Hussain, E., A. Bhaskar, and E. Chung. 2021. "Transit OD Matrix Estimation Using Smartcard Data : Recent Developments and Future Research Challenges [Online]." *Transportation Research Part C*. Elsevier Ltd 125:103044. https://doi.org/10.1016/j.trc.2021.103044.

Jung, J., and K. Sohn. 2017a. "Deep-Learning Architecture to Forecast Destinations of bus Passengers from Entry-Only Smart-Card Data, *IET Intelligent Transport Systems*." *Institution of Engineering and Technology (IET)* 11 (6): 334–339. https://doi.org/10.1049/iet-its.2016.0276.

Kumar, P., A. Khani, and Q. He. 2018. "A Robust Method for Estimating Transit Passenger Trajectories Using Automated Data." *Transportation Research Part C* 95:731–747. https://doi.org/10.1016/j.trc.2018.08.006.

Lahat, D., T. Adali, and C. Jutten. 2015. "Multimodal Data Fusion: An Overview of Methods, Challenges, and Prospects." *Proceedings of the IEEE* 103 (9): 1449–1477. https://doi.org/10.1109/JPROC.2015.2460697.

Lee, S., J. Lee, B. Bae, D. Nam, and S. Cheon. 2021. "Estimating Destination of bus Trips Considering Trip Type Characteristics." *Applied Sciences (Switzerland)*. MDPI 11:21.

Liu, W., Q. Tan, and L. Liu. 2020. "Destination Estimation for Bus Passengers Based on Data Fusion." *Mathematical Problems in Engineering* 2020.

Ma, Xiaolei, Yao-Jan Wu, Yinhai Wang, Feng Chen, and Jianfeng Liu. 2013. "Mining Smart Card Data for Transit Riders' Travel Patterns [Online]." *Transportation Research Part C: Emerging Technologies*. Elsevier Ltd 36:1–12. https://doi.org/10.1016/j.trc.2013.07.010.

Munizaga, Marcela, Flavio Devillaine, Claudio Navarrete, and Diego Silva. 2014. "Validating Travel Behaviour Estimated from Smartcard Data [Online]." *Transportation Research Part C: Emerging Technologies*. Elsevier Ltd 44:70–79. https://doi.org/10.1016/j.trc.2014.03.008.

Munizaga, Marcela A., and Carolina Palma. 2012. "Estimation of a Disaggregate Multimodal Public Transport Origin–Destination Matrix from Passive Smartcard Data from Santiago, Chile." *Transportation Research Part C: Emerging Technologies* 24: 9–18. http://dx.doi.org/10.1016/j.trc.2012.01.007.

Nassir, Neema, Alireza Khani, Sang Gu Lee, Hyunsoo Noh, and Mark Hickman. 2011. "Transit Stop-Level Origin-Destination Estimation Through Use of Transit Schedule and Automated Data Collection System." *Transportation Research Record* 2263 (1): 140–150. https://doi.org/10.3141/2263-16.

Nunes, A. A., T. G. Dias, and J. F. Cunha. 2015. "Passenger Journey Destination Estimation from Automated Fare Collection System Data Using Spatial Validation." *IEEE Transactions on Intelligent Transportation Systems* 17 (1): 133–142. https://doi.org/10.1109/TITS.2015.2464335.

Ordóñez Medina, S. A. 2018. "Inferring Weekly Primary Activity Patterns Using Public Transport Smart Card Data and a Household Travel Survey." *Travel Behaviour and Society* 12:93–101. https://doi.org/10.1016/j.tbs.2016.11.005.

Sari Aslam, N. 2022. *Inferring Trip Purposes from Travel Smart Card Data*. Doctoral thesis. London: UCL.

Sari Aslam, N., T. Cheng, and J. Cheshire. 2019. "A High-Precision Heuristic Model to Detect Home and Work Locations from Smart Card Data [Online]." *Geo-spatial Information Science*. Taylor and Francis 22 (1): 1–11. https://doi.org/10.1080/10095020.2018.1545884.

Sari Aslam, Nilufer, Di Zhu, Tao Cheng, Mohamed R. Ibrahim, and Yang Zhang. 2020. "Semantic Enrichment of Secondary Activities Using Smart Card Data and Point of Interests : A Case Study in London [Online]." *Annals of GIS*. Taylor and Francis 27 (1): 1–13. https://doi.org/10.1080/19475683.2020.1783359.

Shenzhen Metro Group Co. 2022. *Basic Fare and Discount Policy* [Online]. Doi: https://www.szmc.net/szmc_en/Tickets_and_Fares/Basic_Fare_and_Discount_Policy/ [Accessed on: 28-11-2022].

Tao, Sui, David Rohde, and Jonathan Corcoran. 2014. "Examining the Spatial–Temporal Dynamics of Bus Passenger Travel Behaviour Using Smart Card Data and the Flow-Comap." *Journal of Transport Geography* 41: 21–36. http://dx.doi.org/10.1016/j.jtrangeo.2014.08.006.

Trépanier, M., and R. Chapleau. 2006. "Destination Estimation from Public Transport Smartcard Data." *IFAC Proceedings Volumes (IFAC-PapersOnline)* 12 (PART 1).

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*.

Wang, W., J. P. Attanucci, and N. H. M. Wilson. 2011. "Bus Passenger Origin-Destination Estimation and Related Analyses Using Automated Data Collection Systems." *Journal of Public Transportation* 14 (4): 131–150. http://dx.doi.org/10.5038/2375-0901.

Yan, F., C. Yang, and S. V. Ukkusuri. 2019. "Alighting Stop Determination Using two-Step Algorithms in bus Transit Systems [Online]." *Transportmetrica A: Transport Science* 15 (2): 1522–1542. https://doi.org/10.1080/23249935.2019.1615578.

Yang, Yuanxuan, Alison Heppenstall, Andy Turner, and Alexis 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. https://doi.org/10.3390/ijgi8060271.

Yu, C., and Z. C. He. 2017. "Analysing the Spatial-Temporal Characteristics of bus Travel Demand Using the Heat map." *Journal of Transport Geography*. Elsevier Ltd 58:247–255. https://doi.org/10.1016/j.jtrangeo.2016.11.009.

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. https://doi.org/10.1111/j.1467-8667.2007.00494.x.

Zou, Q., X. Yao, P. Zhao, H. Wei, and H. Ren. 2016. "Detecting Home Location and Trip Purposes for Cardholders by Mining Smart Card Transaction Data in Beijing Subway." *Transportation*. Springer US 45:919–944.