Extracting bus transit boarding stop information using smart card transaction data

Zhen Chen[1] · Wei Fan[1]

Received: 1 February 2018 / Revised: 22 May 2018 / Accepted: 27 May 2018 / Published online: 8 June 2018

Abstract The smart card-based automated fare collection (AFC) system has become the main method for collecting urban bus and rail transit fares in many cities worldwide. Such smart card technologies provide new opportunities for transportation data collection since the transaction data obtained through AFC system contains a significant amount of archived information which can be gathered and leveraged to help estimate public transit origin–destination matrices. Boarding location detection is an important step particularly when there is no automatic vehicle location (AVL) system or GPS information in the database in some cases. With the analysis of raw data without AVL information in this paper, an algorithm for trip direction detection is built and the directions for any bus in operation can be confirmed. The transaction interval between each adjacent record will also be analyzed to detect the boarding clusters for all trips in sequence. Boarding stops will then be distributed with the help of route information and operation schedules. Finally, the feasibility and practicality of the methodology are tested using the bus transit smart card data collected in Guangzhou, China.

Keywords Transit smart card · Automated fare collection · Boarding location inference

✉ Wei Fan
wfan7@uncc.edu

Zhen Chen
zchen28@uncc.edu

[1] USDOT Center for Advanced Multimodal Mobility Solutions and Education (CAMMSE), Department of Civil and Environmental Engineering, The University of North Carolina at Charlotte, EPIC Building Room 3366, 9201, University City Boulevard, Charlotte, NC 28223-0001, USA

1 Introduction

As smart card (SC) fare system has been widely implemented in the world today, smart card data (SCD) plays an important role in the regional transportation system management. SCD can act as the data source to replace traditional travel surveys in many respects. Specifically, the transit rider origin–destination (OD) matrix can potentially be extracted from the SC transaction database. Passengers' OD matrix estimation is a major part of any transportation planning study. With the consideration of inconvenience of traditional travel survey methods, SCD-based research has become more and more popular recently. The applications of SCD study are very broad. For example, Cui [1] utilized iterative proportional fitting and maximum likelihood estimation for both single route and network level OD estimation. Devillaine et al. [2] presented a methodology to categorize trips by different purposes after the detection and estimation of the locations of destination, trip time, activity duration and card type by utilizing smart card databases. Kusakabe and Asakura [3] developed a data fusion method to describe and add behavioral attributes to the smart card data by using survey-based data. By applying the probability distribution of behavior attributes to naive Bayes classifier, the attribute of trip purposes ('commuting to work'; 'commuting to school'; 'leisure'; 'business'; 'returning home') is added to each trip in the smart card database. Long et al. [4] conducted an analysis based on the travel time characteristics of four major passenger groups ('early birds,' 'night owls,' 'tireless itinerants' and 'recurring itinerants') with the help of household survey data. Kieu et al. [5] presented a study utilizing three levels of density-based spatial clustering of application with noise (DBSCAN) algorithm for mining transit users' travel regularity (spatial regular OD and temporal habitual traveling time) from SCD. Ma et al. [6] developed an effective data-mining methodology utilizing Beijing transit data and also applied DBSCAN algorithm to analyze and detect transit riders' historical travel patterns.

Although smart card fare systems are essentially different in each city and the relevant data components are unique in each case, in general, the first step of SCD-based OD estimation is to gather the boarding and alighting information. It is also important to note that in many transit systems, the smart card system is entry-swipe only which may not even have the information about boarding stops/locations. As such, different research methods need to be developed and applied for building up the OD matrix. In this regard, a number of SCD studies have been conducted to infer the OD matrix based on two important and yet reasonable assumptions as pointed out by Barry et al. [7]: (1) 'Most riders would like to return the destination station of their last trip to begin next trip; and (2) The destination of most riders at the end of day is the origin stop where they begin their first trip of the day.' Based on these assumptions, inter-transaction time (ITT) was considered as a key element in transfer activities detection and OD estimation by Alsger et al. [8]. As for the buffer zone range calibration, different allowable transfer times (i.e., the threshold of ITT) were assumed which ranged from 30 to 90 min [6, 9–11].

As mentioned, the trip origin detection is usually the first step of OD matrix inference when working with incomplete database. Association of the GPS data with smart card transaction data is one of the methodologies that can be used to infer boarding stops. Wang et al. [12] inferred the trip origins with the help of both transaction record timestamp and AVL data, and achieved some results with high accuracy. To solve the origin inference problem without the AVL system, Barry et al. [13] developed an approach to identifying the specific boarding stops utilizing scheduled run times to estimate the location of a bus along its route at the time of the AFC transaction. The challenge associated with the location estimation in their study was that the transaction times of MetroCard system were truncated to 6-min intervals. To detect the approximate bus boarding locations, they used scheduled run time between stops along the route and adjusted the results with the transfer information obtained from SCD.

Station labeling algorithm was used to assign each record of a bus route in a sequence. Song [14] utilized a fixed time threshold with an adjustment element to distinguish the boarding records at different stops with the time difference between adjacent records. This is the basic idea of boarding stops location detection. However, there are two problems in this algorithm that may contribute to errors. As known, the operation times between different stops can be different due to varying distances, number of changing boarding passengers and different signalized intersections involved. The specific time threshold between each pair of stops may result in the inaccuracy of stops labeling. Another problem is the consideration of the bus status at each terminal. One cannot be certain as to how long the bus stay will be in the terminal. Hence, using the only criterion 'change transaction interval larger than bus dispatching time at terminus to detect direction change' may not be accurate enough.

With only the transaction time being available, a few research efforts also attempted to adopt a hierarchical clustering method to classify the source data based on the time interval between two adjacent transaction records [15]. The reference bus operation time was based on the fixed schedule and the criterion used was based on the difference between timestamp of the cluster and the arrival time of each stop on the schedule. However, it was very difficult to solve the problem since the clusters could not be built without direction information.

Another approach was presented by Ma et al. [16], in which a Markov chain-based Bayesian decision tree algorithm was developed and utilized to extract passengers' origin information from the Beijing flat-bus AFC database. The first task was to introduce how to cluster transaction data by taking the following two steps: (1) Sort transaction timestamps in ascending order; and (2) Include the record as another cluster if the time difference between itself and the previous record was larger than 60 s in which 30 min was set as the time threshold for a different trip. However, since the irregularity of the sample database and the time threshold configuration were a big issue, direction identification task was not described in detail in this research. In short, the following Table 1 part (a) provides a summary of the existing smart card research efforts for OD estimation and Table 1 part (b) gives a review summary of the existing smart card data research for origin inference.

The purpose of this research is to develop a systematic approach to illustrating how passenger boarding information can be mined only from the raw data derived from smart card-based automated fare collection (AFC) system, without the reference of GPS related data, AVL system, travel survey and other supporting database. In other words, there is only transaction time recorded, and no boarding stops locations and no bus operation direction information are available in this research. Advanced methods are developed to acquire the bus transit operation direction information, passengers boarding time information and the specific boarding stops information for each record. The methodology developed is tested via Guangzhou bus transit smart card data. It is important to note that effectively extracting boarding location information can greatly help build up the trip chain and also determine the destination information. The OD matrix can then be

Table 1 Summary of existing smart card research efforts

Authors	Year	Data type	Data source	Case study location	Methodology	Achievements
(a) Smart card research efforts for OD estimation						
Barry et al.	2002	Card ID; bus ID; boarding transaction time; exist counts	SCD, survey data	NYC, US	Trip chain built up by sequence	Estimated OD matrix
Trépanier et al. [17]	2007	Card ID; boarding transaction time; boarding stops; card type	SCD, survey data	Gatineau, Canada	Trip chain method; comparison with household survey data	Estimated OD matrix
Cui	2006	Card ID; route ID; boarding transaction time; last route ID; boarding and alighting counts at stops	SCD, AVL	Chicago, US	Iterative Proportional Fitting (IPF) and Maximum Likelihood Estimation (MLE)	Estimated OD matrix at the segment level (3–5 combined stops)
Zhang et al. [19]	2007	Card ID; boarding transaction time; driver ID; route ID; on board survey data	SCD, survey data	Changchun, China	Comparison of records with on bus survey data	Estimated OD matrix
Seaborn [21]	2009	Card ID; mode; journey stage sequence number; start information; end information; date	SCD	London, UK	Iterative method; trip chain method	Estimated OD matrix; estimated route connectivity information
Wang [18]	2010	Card ID; route ID; Boarding transaction time; boarding stops; Alighting transaction time; alighting stops; scheduled departure time; actual departure time; GIS data	SCD, AVL	London, UK	Matching records with AVL data; trip chain method, compared the results with survey data	Estimated OD matrix
Nassir et al.	2011	Card ID; Card type; Boarding transaction time; Boarding stops; User type; Bus ID	SCD, AVL	Minneapolis, US	Matching records with AVL data; Sensitivity analyses	Estimated alighting stop information
Kieu et al.	2013	Card ID; route ID; boarding transaction time; boarding stops; alighting transaction time; alighting stops; Trip direction	SCD, survey data	Brisbane city, Australia	Travel itineraries reconstruction; DBSCAN algorithm	Revealed SC holders' historical travel patterns
Devillaine et al.	2013	Card ID; card Type; boarding transaction time; boarding stops; bus ID	SCD	Gatineau, Canada	Trip chain method;	Identified transit users' trip characteristics
Alsger et al.	2014	Card ID; route ID; boarding transaction time; boarding stops; alighting transaction time; alighting stops	SCD	South East Queensland, Australia	Trip chain method; ITT analysis	Identified best ITT time threshold
Ma et al.	2014	Card ID; route ID; boarding transaction time; boarding stops; alighting transaction time; alighting stops	SCD	Beijing, China	Trip chain method; DBSCAN algorithm; K-means clustering	Revealed SC holders' historical travel patterns
Kusakabe and Asakura	2014	Trip ID; boarding information; alighting information; Trip purpose; card ID	SCD, Survey data	NA	Naive Bayes probabilistic model	Enhanced understanding of travelers' behavior
Chang and Zhao [20]	2016	Card ID; boarding transaction time; boarding stops; route ID; Bus ID	SCD	Guangzhou, China	Trip chain method; DBSCAN clustering	Revealed SC holders' travel patterns
(b) Smart card research efforts for origin inference						
Barry et al.	2009	Card ID; bus ID; Boarding transaction time	SCD, schedule data	NYC, US	Matching records with schedule trip chain method	Extracted passengers' boarding stop information
Yu et al.	2009	Card ID; route ID; boarding transaction time; bus ID; bus operation data	SCD	NA	Hierarchical clustering method	Extracted boarding stop information

Table 1 continued

Authors	Year	Data type	Data source	Case study location	Methodology	Achievements
Wang et al.	2011	Card ID; route ID; boarding transaction time; bus ID; GPS data; bus operation data	SCD, AVL data	London, UK	Matching records with schedule and GPS data	Extracted boarding stop information
Ma et al.	2012	Card ID; boarding transaction time; bus ID; route ID	SCD	Beijing, China	Transaction data clustering; Markov chain-based Bayesian decision tree algorithm	Extracted transit passengers' origin
Song	2016	Card ID; route ID; boarding transaction time; bus ID	SCD, schedule data	Chengdu, China	Time labeling method; greedy algorithm	Extracted boarding stop information

developed and used to help decision-makers plan, design, operate, and manage a more efficient public transit system. The contributions of this study include developing a systematic methodology which can be applied to most of the SCD systems and generating the OD estimation results which can reveal bus transit users' travel pattern based on the collected SCD, both of which will greatly help transit operators and planning organizations develop objective and effective data-supported policies.

The remainder of this paper is organized as follows: Sect. 2 presents the methodology. The structure and potential problem of the smart card data are discussed. The direction identification algorithm is then described. This is followed by the boarding activities clustering and boarding stops identification. Section 3 discusses the numerical results of this study in detail. Finally, conclusions of the study are made, and the future research directions are also given in Sect. 4.

2 Methodology

This section describes how to estimate boarding location information in detail. It first describes how to supplement the data when there is missing information about the transit directions and boarding clusters. Second, this section discusses the relationship between records in the sample database and presents the description of the boarding location inference methodology.

2.1 Basic data information

The study utilizes Guangzhou, China, transit system smart card database for transit OD estimation and trip purposes analyses. A transaction record is generated each time a passenger boards. Data on 4 bus routes in 5 days are used for analyses which included 100,000 transactions.

The smart card records came from the bus transit smart card system. The fare collection system used entry-swipe only. The information was gained from smart card reader and then transferred to data management center. Since the system was not integrated with the AVL system, the data recorded did not have the stop ID as to where the passengers paid their fares and rather, it had a transaction time record and information about the transit route. Neither the alighting information was recorded.

A sample of the transaction records of the smart card system is shown in Table 2 which contains the following information:

- *Route* This field indicates which route the transit record belongs to.
- *Card ID* The card ID is a 16-digit number which uniquely identifies a smart card. This field is critical as it allows selection of transit trips made by a particular passenger. Since the system allows multiple users to share the same smart card, additional processing will be required to eliminate the redundant records in that case.
- *Card type* The card type in this study includes general card, student card, employee card, disable card and senior card. The different categories of the card can help detect transit users' activity characteristics in an effective manner.
- *Bus ID* This field indicates on which bus the transaction occurred. The bus ID information can help infer the direction of the transit.
- *Transaction time* Transaction time field contains accurate timestamp information at the 'second' level. This field is necessary to sort the transit trips (transactions) of a passenger in a sequential order.

In order to ensure data quality, the raw data are processed in which the data cleaning step is conducted first. The major problem is related to the consecutive swiping activities. Although group travelers have a preference to

Table 2 A sample of the smart card raw data

Route	Card ID	Card type	Bus ID	Transaction time
Route A	‘9999995338410746	General	‘60750014	‘20150101011816
Route A	‘9999990069526128	General	‘60750014	‘20150101002722
Route A	‘9999998626055540	Student	‘60750014	‘20150101002658

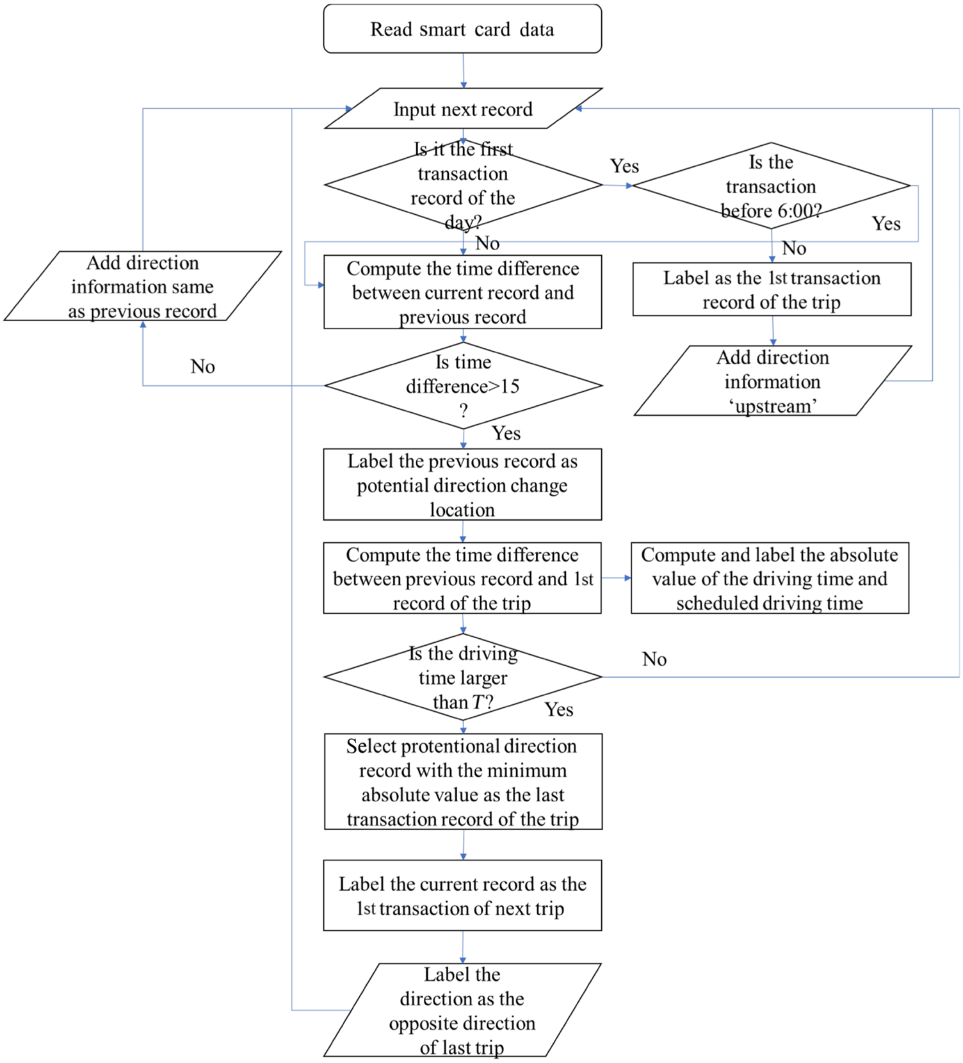

Read smart card data
Input next record
Is it the first transaction record of the day?
Yes
Is the transaction before 6:00?
Yes
No
No
Label as the 1st transaction record of the trip
Add direction information ‘upstream’
Compute the time difference between current record and previous record
Add direction information same as previous record
No
Is time difference>15 ?
Yes
Label the previous record as potential direction change location
Compute the time difference between previous record and 1st record of the trip
Compute and label the absolute value of the driving time and scheduled driving time
Is the driving time larger than T?
No
Yes
Select protentional direction record with the minimum absolute value as the last transaction record of the trip
Label the current record as the 1st transaction of next trip
Label the direction as the opposite direction of last trip

Fig. 1 Trip direction identification algorithm

swipe their cards continuously, the records of same card ID with hundreds of continuous swipes seem to involve some types of errors. Therefore, the records with more than three consecutive swiping activities are removed from the raw data. Specifically, the data cleaning step is presented below:

Step 1 Sort the data by card ID.

Step 2 Label the records of the same card ID occurred during same trip as consecutive swipe records.

Step 3 Remove the records of the same card ID with more than three swiping activities.

2.2 Route information

As there is no available published/fixed schedule of the buses, the operation times between each station are computed based on the operation length and number of intersections between bus stations, as well as the total distance

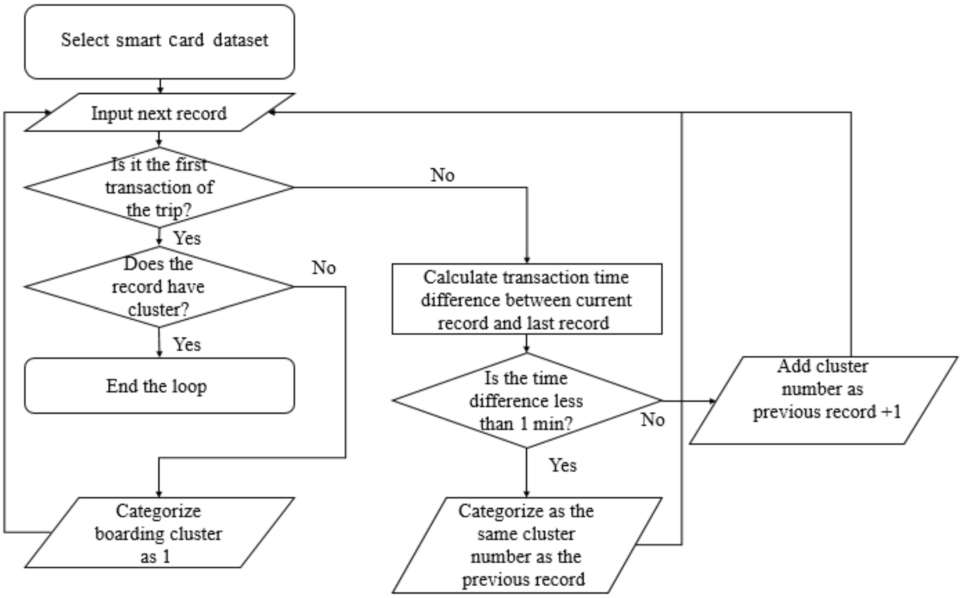

Select smart card dataset
Input next record
Is it the first transaction of the trip?
No
Yes
Does the record have cluster?
No
Yes
End the loop
Calculate transaction time difference between current record and last record
Is the time difference less than 1 min?
No
Yes
Add cluster number as previous record +1
Categorize boarding cluster as 1
Categorize as the same cluster number as the previous record

Fig. 2 Boarding cluster identification algorithm

of the route. The distance between each station can be detected via GIS data of the route. 25 km/h is assumed and utilized as the operation speed of all buses [22]. The travel time between stops can be illustrated as

$$T_{ij} = D_{ij}/V_{\text{average}} + \left(T - D_{\text{total}}/V_{\text{average}}\right) \times U_{ij}, \qquad (1)$$

where D_{ij} stands for the distance between two adjacent stops i and j on the map; V_{average} *is* the average operation speed of the bus (i.e., 25 km/h in this study); T is the total travel time of a whole bus trip on the schedule (i.e., from first station to the terminal station); U_{ij} is the adjustment made to the travel time between stops, which can be different due to different factors including unsignalized intersections, U-turns, roundabouts and traffic signals involved.

2.3 Direction information identification

Since the transaction database only recorded the transaction time and contained no information about the bus direction, the boarding stops cannot be determined by matching transaction times with the estimated transit operation schedule directly. To infer trip direction information, many studies utilized a 30-min time gap as the time threshold of direction identification [16]. Note that such time gap is not the only way which can be used to detect direction change because the time spent by the buses at terminal station is unknown. For example, when the bus is operated during peak hours, it may leave the terminal after only a short break due to the needs to serve high passenger demands. To detect the bus trip direction, it is necessary to study the transaction records of each bus based on the time sequence, and label the direction with the original operation time per trip on the schedule as a reference. In this regard, the direction of the first transaction period is marked. The direction labeling process will use the following criteria:

- The direction will be changed when the time gap between the current transaction and last transaction is more than 30 min.
- The time difference between current record and first record of last transaction sequence is close to the operation time on schedule.
- The records with the transaction time before 6 a.m. on any day will be classified as belonging to the trip series of the previous day.

In short, the trip direction detection algorithm is presented below:

Step 1 Sort the data by route ID/bus ID/transaction date/transaction Time.

Step 2 Label the direction as 1 (upstream) to the first transaction record of the day after 6 a.m.

Step 3 Read and label next transaction record Record based on the time difference between current record j and previous record i as t_{ij} and also label the time difference between current record and the 1st record as t'_j.

Step 4 Label direction information if $t_{ij} < 15$ min and $t'_i < T$ (total operation time for a whole bus trip along this route on the schedule), label the direction of current record same as the previous one. If $t_{ij} > 15$ min, and the t'_i of the previous record $< T$, then label the status of current record as 'hold.' Keep reading, when another $t_{ij} > 15$ min and $t'_i > T$, determine the direction change status based on the value $|t'_i - T|$ and choose the record with the minimum (among all $|t'_i - T|$ values) as the direction change record.

The following flowchart in Fig. 1 illustrates the details of the algorithm.

2.4 Transaction data clustering

As known, many different scenarios could exist in the real world. For example, congestion can occur during peak hours, and there may be no boarding activities at the first and/or last stops. Therefore, it is impossible to guarantee that the bus will be operated according to the fixed schedule. Based on the transaction records, it will be convenient and also more accurate to categorize several transaction records together as boarding clusters. Because several passengers usually board in an intensive period of time and multiple smart card swiping activities will therefore occur at one specific bus stop, the boarding clusters can be labeled based on the time interval between the transactions. The following flowchart in Fig. 2 presents the detailed information about the boarding cluster identification algorithm.

The process contains the following steps: (1) The records are sorted by the sequence of route ID/bus ID/transaction date/transaction time; and (2) The time interval threshold for two consecutive records is 60 s [16]. If the interval is within 60 s, records are categorized as the same boarding cluster; otherwise, the boarding cluster will be changed.

Table 3 shows an example of the clustering results. The number of clusters indicates the boarding activities during a whole trip along a route. It is obvious that there can be no boarding passengers at some stops. For each cluster, timestamps will be added to both the first and last record for the convenience of future analyses.

2.5 Boarding stop information extraction

After the identification of boarding clusters, the specific boarding stops can be inferred based on the difference in timestamps between adjacent boarding clusters. The first boarding cluster is assumed to belong to the first stop (i.e., origin) of the route.

Based on the results of direction identification, the bus operation time of most trips are shorter than the scheduled operation time of a whole bus trip. By examining the data, this can be interpreted as the few passengers' boarding activities occurring during several intermediate (particularly the last few) bus stops of the trip. In addition, traffic delays should also be considered under the situation of congestion during peak hours. Hence, the boarding stop information extraction will use the following rules, which will ensure that different clusters be assigned to different bus stops:

- The boarding stop of 1st boarding cluster is labeled as 'stop 1.'
- The difference between the timestamps of the 1st record of boarding cluster $n + 1$ and last record of boarding cluster n is Δt_{n+1}, the boarding stop ID of boarding cluster n is i; The operation time between station i and $i + 1$ is $a_{i\ (i+1)}$ on the schedule.
- If $\Delta t_{n+1} < a_{i\ (i+1)}$, label the boarding stop as $i + 1$ to the records with boarding cluster $n + 1$.
- If $\Delta t_{n+1} > a_{i\ (i+1)}$, compare Δt_{n+1} and the operation time $a_{i\ (i+2)}$ between station i and $i + 2$ on the schedule……until $a_{i\ (i+\mathrm{k})} > \Delta t_{n+1}$, then label the boarding stop as $i + k - 1$ to the records with boarding cluster $n + 1$.

Figure 3 presents the rules used in the boarding information extraction process.

3 Numerical results

With the application of the algorithm as developed above to process the smart card data collected from Guangzhou, China, the boarding information is identified successfully which is presented in Table 4. There is a total of 100,000 smart card transactions with 98,632 of them being error-free. An example of such errors can be described as the consecutive swiping activity (e.g., transaction records of one specific smart card ID occurred more than three times in a same bus trip). By analyzing the raw data, information gathered out of this study includes:

1. *Direction information* After the direction information is labeled to each record, the time difference between the first and last records can be achieved, which represents the actual travel time from the 1st stop to the last stop with boarding records/activities.
2. *Boarding cluster information* With the help of boarding cluster identification, the transaction records for a whole trip along a route are divided into different cluster groups in order to identify the transactions occurred at different stops. The results of cluster identification can greatly reduce potential errors involved in the boarding stop identification and boarding passenger count estimation.
3. *Boarding stop information* Finally, the results show the estimated boarding location of each transaction records. Additional information about passenger

Table 3 An example of the boarding cluster identification results

Route ID	Transaction ID	Card ID	Card type	Bus ID	Transaction date	Transaction time	Direction	Boarding cluster
Route A	10061	90009698490	Regular card	75	2015/1/1	0:15:05	2	1
Route A	10062	95208436630	Regular card	75	2015/1/1	0:16:54	2	2
Route A	10063	99857026060	Regular card	75	2015/1/1	0:16:55	2	2
Route A	2411	91005456310	Regular card	75	2015/1/1	0:17:31	2	2
Route A	2412	99455451040	Regular card	75	2015/1/1	0:17:33	2	2
Route A	2451	99460138890	Senior card	75	2015/1/1	0:27:04	2	3

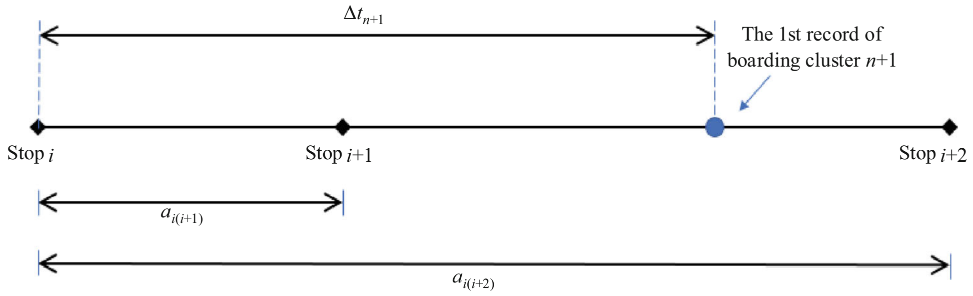


Fig. 3 Boarding information extraction

boarding counts at each stop could be mined by analyzing the SCD with estimated boarding locations.

Table 4 shows an example of the clustering results.

3.1 Comparison of average passenger counts and operation time during each time period

Based on the trip direction identification results, it is possible to calculate the average passenger counts and operation time of each bus trip during each time period. The results with only one boarding cluster in a whole trip are excluded since it will be impossible to calculate the operation time for such records. The passenger boarding counts during each period are also presented in the chart. The passenger counts results indicate that the crest value of boarding activity occurred during different peak hours. During the AM period, the highest volume occurred at 8–9 a.m. and the top three periods were 8–9, 10–11 and 9–10 a.m. During the PM period, the highest volume occurred at 5–6 p.m. and the top three periods are 5–6, 6–7, and 4–5 p.m. The reason behind this could be explained as follows: Most of government institution and enterprises begin their work around 9 a.m. and finish their work around 6 p.m. Therefore, the travel patterns of citizens in Guangzhou follow exactly the same. This result is also consistent with previous studies (e.g., [23, 24]).

Furthermore, the two charts in Fig. 4 also indicate that the operation time will increase as the passenger boarding counts increase.

3.2 Frequency of passengers' boarding activities at each stop

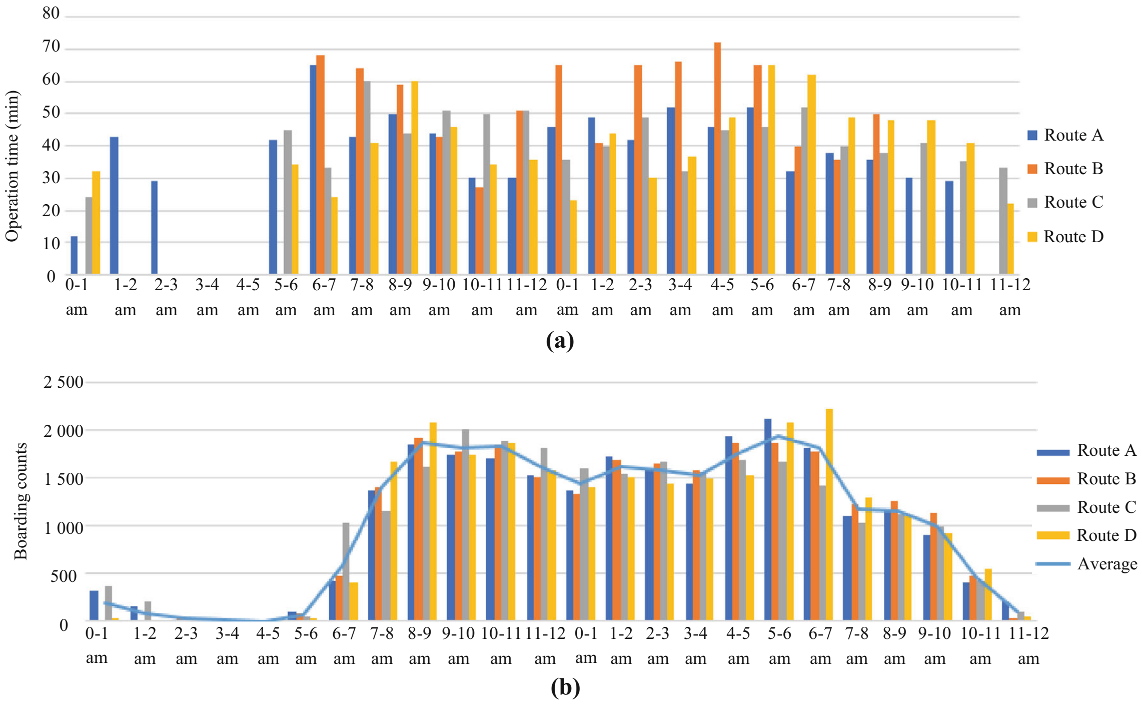
Based on the boarding stop identification results of each record, it is also possible to estimate the frequency of boarding activities at each stop. As shown in the results in Fig. 5, most boarding activities occur at the first several stops, and the passenger boarding counts decrease as the bus stops get closer to the terminal. The passenger boarding activities rarely occur at last several stops. The reason for high passenger volume occurrence at 1st stop could be explained as the influence of the assumption of labeling the boarding stop of 1st boarding cluster as stop 1. The decreasing trend at last several stops is also consistent with previous studies (e.g., [23]). That clearly indicates both the travel habits of local passengers and the passenger boarding location characteristics from the traffic network point of view.

4 Conclusions

It is very challenging for one to conduct the OD estimation if there is no boarding stop information recorded by the AFC system. This research aims to develop a methodology to extract the boarding location information using the available transaction records with only basic route information, transaction time and few transfer activities, but without GPS, passenger counts, and survey data. To reduce

Table 4 An example of the boarding stop identification results

Route ID	Transaction ID	Card ID	Card type	Bus ID	Transaction date	Transaction time	Direction	Boarding cluster	Boarding stop ID
Route A	10044	99852761030	Regular card	75	2015/1/1	0:14:44	2 (downstream)	1	1
Route A	10061	90009698490	Regular card	75	2015/1/1	0:15:05	2 (downstream)	1	1
Route A	10062	95208436630	Regular card	75	2015/1/1	0:16:54	2 (downstream)	2	2
Route A	10063	99857026060	Regular card	75	2015/1/1	0:16:55	2 (downstream)	2	2
Route A	2411	91005456310	Regular card	75	2015/1/1	0:17:31	2 (downstream)	2	2
Route A	2412	99455451040	Regular card	75	2015/1/1	0:17:33	2 (downstream)	2	2
Route A	2451	99460138890	Senior card	75	2015/1/1	0:27:04	2 (downstream)	3	5



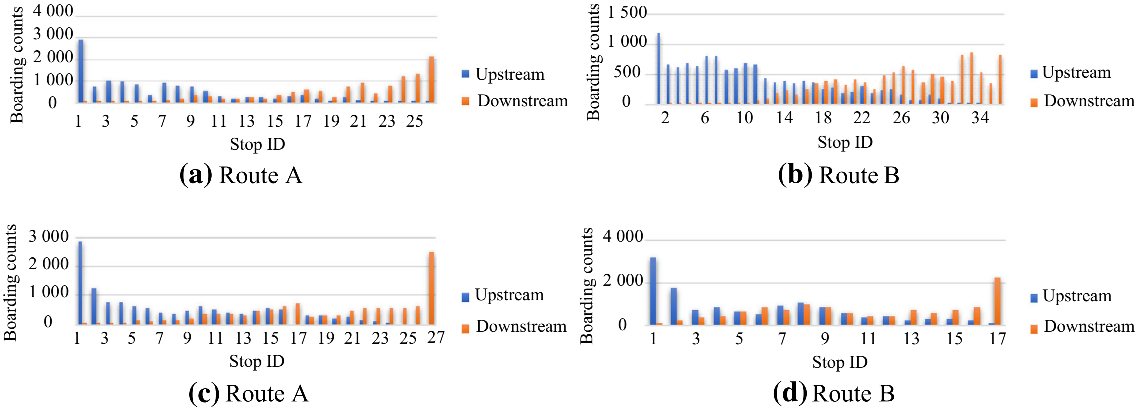
(a)

(b)

Fig. 4 Boarding counts and operation time during each time period. **a** Average bus operation time during each time period, **b** passenger boarding counts during each time period

the errors, the algorithm processes the data in the order of identification of boarding direction, boarding clusters and boarding stops. The boarding stop information is successfully extracted. Based on the acquired boarding information, the results of passenger counts are also derived in which the crest values of boarding activity occur during different peak hours. During the AM period, the highest volume occurs at 8–9 a.m. and the top three periods are 8–9, 10–11 and 9–10 a.m. During the PM period, the highest volume occurs at 5–6 p.m. and the top three periods are 5–6, 6–7, and 4–5 p.m. Most boarding activities occur at the first several stops, and the passenger boarding counts decrease as the bus stops get closer to the terminal.

However, several improvements could be made in the future, which include the consideration of variable speeds of the buses as one of the potentially influencing factors to calculate the operation time between bus stops. The direction changing time threshold of 30 min could also be further investigated as the fixed time threshold is utilized in this study. It is well noted that this time threshold could be changed during different times of day or days of week if the more transit data are available in the future.



(a) Route A **(b)** Route B

(c) Route A **(d)** Route B

Fig. 5 Boarding counts at each stop

The methodology and results of this study can be helpful for the OD estimation related work in the real world. However, with the limited amount of SCD, the trip chain building up process and destination identification framework are not discussed in this study. In the future, the transfer activities could be mined from the database if it contains information for more routes and longer periods. The cluster analysis can also be conducted to reveal passengers' travel patterns with the help of survey data, card type, and land use data. Furthermore, the alighting stop information can also be extracted and used to estimate the relevant OD matrix.

Acknowledgements The authors want to express their deepest gratitude to The United States Department of Transportation, University Transportation Center through the Center for Advanced Multimodal Mobility Solutions and Education (CAMMSE) at The University of North Carolina at Charlotte (Grant Number: 69A3551747133) for sponsoring this research project entitled 'estimation of origin–destination matrix and identification of user activities using public transit smart card data.' The authors also would like to thank the www.openits.cn Web site for providing the data for this research.

References

1. Cui A (2006) Bus passenger origin–destination matrix estimation using automated data collection systems, Dissertation, Massachusetts Institute of Technology
2. Devillaine F, Munizaga M, Trépanier M (2012) Detection of activities of public transport users by analyzing smart card data. Transp Res Rec J Transp Res Board 2276:48–55
3. Kusakabe T, Asakura Y (2014) Behavioural data mining of transit smart card data: a data fusion approach. Transp Res Part C Emerg Technol 46:179–191
4. Long Y, Liu X, Zhou J (2016) Early birds, night owls, and tireless/recurring itinerants: an exploratory analysis of extreme transit behaviors in Beijing, China. Habitat Int 57:223–232
5. Kieu L, Bhaskar A, Chung E (2013) Mining temporal and spatial travel regularity for transit planning. In: Proceeding of 36th Australasian Transport Research Forum (ATRF), Brisbane, Queensland, Australia
6. Ma X, Wu Y, Chen F, Liu J (2013) Mining smart card data for transit riders' travel patterns. Transp Res Part C Emerg Technol 36:1–12
7. Barry J, Newhouser R, Rahbee A, Sayeda S (2002) Origin and destination estimation in New York City with automated fare system data. Transp Res Rec J Transp Res Board 1817:183–187
8. Alsger A, Mesbah M, Ferreira L, Safi H (2015) Public transport origin–destination estimation using smart card fare data. In: Proceeding of transportation research board 94th annual meeting, Washington, DC, United States
9. Nassir N, Khani A, Lee S, Noh H, Hickman M (2011) Transit stop-level origin–destination estimation through use of transit schedule and automated data collection system. Transp Res Rec J Transp Res Board 2263:140–150
10. Bagchi M, White P (2004) What role for smart-card data from bus systems? Munic Eng 157(1):39–46
11. Hofmann M, O'Mahony M (2005) Transfer journey identification and analyses from electronic fare collection data. In: Intelligent transportation systems, 2005 proceedings. IEEE, pp 34–39
12. Wang W, Attanucci J, Wilson N (2011) Bus passenger origin–destination estimation and related analyses using automated data collection systems. J Public Transp 14(4):131–150
13. Barry J, Freimer R, Slavin H (2009) Use of entry-only automatic fare collection data to estimate linked transit trips in New York City. Transp Res Rec J Transp Res Board 2112:53–61
14. Song Z (2016) Research on large scale OD matrix estimation method based on bus IC card data. Appl Res Comput 33(7):2007–2013
15. Yu Y, Deng T, Xiao Y (2009) A novel method of confirming the boarding station of bus holders. J Chongqing Jiaotong University (Nat Sci) 28(1):121–125
16. Ma X, Wang Y, Chen F, Liu J (2012) Transit smart card data mining for passenger origin information extraction. J Zhejiang Univ Sci C 13(10):750–760

17. Trépanier M, Tranchant N, Chapleau R (2007) Individual trip destination estimation in a transit smart card automated fare collection system. J Intell Transp Syst 11(1):1–14
18. Wang W (2010) Bus passenger origin–destination estimation and travel behavior using automated data collection systems in London, UK. Dissertation, Massachusetts Institute of Technology
19. Zhang L, Zhao S, Zhu Y, Zhu Z (2007) Study on the method of constructing bus stops OD matrix based on IC card data. In: Proceeding of wireless communications, networking and mobile computing, WiCom 2007. International IEEE conference, Shanghai, China
20. Chang Y, Zhao C (2016) Travel pattern recognition using smart card data in public transit. Int J Emerg Eng Res Technol 4(7):6–13
21. Seaborn C (2009) Smart card data for multi-modal network planning in London: five case studies. In: European transport conference, Leeuwenhorst, Netherlands
22. Tantiyanugulchai S, Bertini R (2003) Arterial performance measurement using transit buses as probe vehicles. In: Intelligent transportation systems, 2003. Proceedings. IEEE, pp 102–107
23. Liu L, Hou A, Biderman A, Ratti C, Chen J (2009). Understanding individual and collective mobility patterns from smart card records: a case study in Shenzhen. In: 12th International IEEE conference on intelligent transportation systems, St. Louis, MO, USA
24. Tao S, Rohde D, Corcoran J (2014) Examining the spatial–temporal dynamics of bus passenger travel behaviour using smart card data and the flow-comap. J Transp Geogr 41:21–36