Hindawi
Journal of Advanced Transportation
Volume 2020, Article ID 5740521, 13 pages
https://doi.org/10.1155/2020/5740521

Research Article

A Method for Bus OD Matrix Estimation Using Multisource Data

Di Huang,[1] **Jun Yu**,[1] **Shiyu Shen**,[2] **Zhekang Li**,[1] **Luyun Zhao**,[2] **and Cheng Gong**[2]

[1]*Jiangsu Key Laboratory of Urban ITS, Jiangsu Province Collaborative Innovation Center of Modern Urban Traffic Technologies, School of Transportation, Southeast University, Nanjing 211189, China*
[2]*Didi Chuxing, Beijing 100085, China*

Correspondence should be addressed to Di Huang; dhuang2@seu.edu.cn

Received 14 December 2019; Revised 24 January 2020; Accepted 3 February 2020; Published 21 March 2020

Guest Editor: Tao Liu

The automated fare collection (AFC) system has gained increasing popularity among transit systems worldwide. The AFC system is usually an entry-only system that only records the serial number of the smart card and the transaction time of each use. Neither the AFC data nor the bus global positioning system (GPS) could reveal the passenger's alighting information, namely, alighting time and station. Hence, the station-to-station origin-destination (OD) trip information cannot be obtained directly from the available data sources. To address this problem, this paper proposes a methodology that estimates the OD matrix by using smart card and GPS data. In this paper, the characteristics of the basic data sources are first analyzed, based on which the bus arrival time is generated using the density-based clustering algorithm and a time correction strategy, based on which the passenger's boarding station is identified. The alighting stations are inferred based on the characteristics of bus trip chaining, which could identify over 80% of the alighting stations on average. Finally, the proposed methodology is verified by a comprehensive field survey in Suzhou, China, with 100% sample rate.

1. Introduction

The bus origin-destination (OD) matrix estimation is one of the most fundamental steps for urban transit planning, operation, and management strategies, such as passenger demand forecasting [1, 2], network design [3], transit pricing [4, 5], scheduling [6, 7], and bus operation [8, 9]. Traditionally, the acquisition of OD trip information as well as the analysis of passenger's spatial-temporal travel behavior relies largely on travel survey data, which has the drawbacks of small sample size, high costs, and being time consuming [10–14]. With the help of advanced automatic fare collection (AFC) system, the smart card data has received increasing interest as a new and reliable data source for the collection and analysis of passenger trip information [15]. The statistical performance of smart card data outperforms that of the field survey data for providing comprehensive spatial-temporal information about the transit system and capturing the dynamics of passenger trips [16, 17]. The smart card data could also detect and correct the potential sources of sampling bias [18].

There has been some work in mining the smart card data and other types of data collected during the bus operation, such as bus global positioning system (GPS) data and WiFi/Bluetooth probes data. Pelletier et al. [19] have conducted an extensive review of the cutting-edge technique of smart card data mining in public transit system. Early studies on the smart card data analysis mainly focus on the statistical analysis of urban public transport passenger flow. For example, both Kiyohiro et al. [20] and Ouyang et al. [21] use smart card data to analyze the travel pattern of urban residents. Ebadi et al. [22] utilize smart card data to construct students' activity-mobility trajectories in time-space dimension. Ji et al. [23] propose a Bayesian model to estimate trip-level OD flow matrices utilizing the data collected by Wi-Fi sensors and boarding data provided by automatic vehicle location (AVL) systems. However, compared to the smart card data, the stability of Wi-Fi sensors and the reliability of collected data are relatively low. Ma et al. [24] investigate the smart card transactions and GPS data. The results show relatively high accuracy of the passenger's origin information. Ma et al. [25] continue to mine the smart card data and

propose an efficient method to identify the passenger's travel patterns using passenger's historical travel data.

It has been widely recognized that the application of data fusion techniques which consider various correlated data sources can help to improve the accuracy of the estimation of OD trip matrix [26–29]. In such data fusion process, the following two issues have to be addressed: (1) the smart card data contains only transaction-related information, which is useful only when it can be matched with the vehicle's operational data while the boarding and alighting locations are not recorded [30]; (2) the information recorded by the smart card system might not be matched perfectly with that data of the GPS system. To identify the boarding station, two kinds of techniques are mainly adopted, namely, clustering analysis and GPS data matching. Specifically, the swiping time of ID card is first clustered and then matched with the AVL data to identify bus stops. Many efforts have been devoted for combining the smart card data with the bus dispatching information. Zhang et al. [31] identify the commuting trips of passengers by using their smart card data and use the clustering analysis method to identify commuting passengers' boarding stations according to the time recorded in the smart card data. Farzin [32] uses the GPS time matching to identify the boarding location through the correlation matching of bus IC card swiping information, GPS positioning information, and bus station locations in São Paulo, Brazil.

The identification of the alighting station is more challenging because of the absence of available alighting location information [19]. Two kinds of methods have been proposed to estimate passenger's alighting station: aggregate and disaggregate methods. In the aggregate method, passengers are assumed to alight from the bus according to a specific probability distribution with respect to the travel distance and station attractiveness [33]. To obtain more reliable estimations, a large number of studies focus on the disaggregate method based on the passenger's trip chain by combining the smart card data with other sources of data, e.g., AVL data, which provides the detailed trip and corresponding geographic information.

The trip chain is generated to describe the same individual's daily travel which is composed of several bus trips with the assumption that the passenger's boarding station of a trip is close to the alighting station of the last trip [34]. Wang et al. [35] apply the trip chaining to obtain the OD information using smart card transactions and AVL data in London. It is also the first attempt to validate the results with manual passenger survey data. However, the sample rate of some bus lines is about 60%. Barry et al. [13] infer the alighting station based on the data collected by AFC system in New York, USA. Two specific assumptions are made to define the individual's trip chain: (1) a large proportion of passengers returns to the previous station to start the next trip and (2) passengers would begin their first trip of the day at the station where they end their trip of the previous day. Zhao [36] examines the rail-to-bus trip chain by integrating the AFC and AVL data, as well as the GIS technology. It was demonstrated that the passenger's travel behavior can be obtained rigorously by using the advanced automated data collection system. Seaborn et al. [37] use the smart card fare payment data to investigate the passenger's multimodal trips in London. They illustrate that passengers between intersecting routes can be identified by the smart card data. Cui [38] first estimates the OD matrix of a single route and then extends it to the network-level OD matrix estimation. A case study of a full-size network of Chicago is conducted. The results show that the maximum likelihood estimation is suitable for the estimation of the single route OD matrix, while the proportional distribution method is recommended for the estimation of the transfer flow OD matrix in the network level. Lu et al. [39] use the AFC data from Beijing metro based on the trip chaining method and k-means clustering method to infer the actual destination, which could also be applied to the bus passengers' alighting station inference based on the smart card data. In sum, the above-mentioned studies on the passenger's trip chain are based on two assumptions: (1) the closest stop and (2) the daily symmetry rules. The identification rate of the passenger's alighting station using the trip chain method is between 67% and 71% [33, 36, 38].

In literature, bus OD matrix can be estimated in three ways, i.e., field survey, analysis of smart card data only, and the fusion of multisource data including smart card data and bus GPS data. Although existing literature has extensively studied the characteristics of different types of data, dealing with the heterogeneity and systematic error of multisource data is still an open question. Specifically, the gaps of current studies are summarized as follows: (1) the fusion of multisource data does not allow for the internal correlation in both spatial and temporal dimensions; (2) the inference of alighting stations is accurate on an aggregate level but inaccurate on an individual level; and (3) the estimated results cannot be verified comprehensively because of the absence of real data and inadequate sample rate through field experiments.

To sum up, the contribution of this study is threefold. First, this paper proposes a spatiotemporal correction method for smart card data, bus GPS data, and bus station data to address incompleteness and complexity issues of multisource data. A density-based clustering method is applied to correct the difference between the trajectories recorded by the GPS devices and the bus station data. Second, the bus arrival time is obtained through data clustering technique in the spatial dimension. An additional correction algorithm in the temporal dimension is proposed to calibrate the timestamps in smart card data to match the bus arrival time. To identify the alighting station in the individual level, this paper divides the trips into the chained dataset and unchained dataset and infers the alighting stations for individuals for the chained dataset, through which the identification rate of the alighting station can be improved by about 10%. Third, the effectiveness and feasibility of the proposed method are verified on the data collected by a large-scale field survey in Suzhou, China.

The rest of this paper is orgnized as follows.The problem statement and data utilized in this paper are introduced in Section 2. Section 3 discusses the method for identifying boarding stations by incorporating the heterogeneity in multisource data. Section 4 presents the categories of trips and the trip-chaining-based inference method for alighting

Table 1: Structure of smart card data.

Field	Field name	Description
RECORD_ID	Record number	Unique identification of a smart card record
CARD_UUID	IC card number	Unique identification of smart card
SWIPED_TYPE	Card type	General card/senior card/student card
SWIPED_TIME	Swiping time	2018-03-01 13:17:54
LINE_UUID	Line number	Bus line number
BUS_UUID	Vehicle number	Bus number
FARE	Fare	Bus fare

Table 2: Structure of bus GPS data.

Field	Field name	Description
RECORD_ID	Record number	Unique identification of a GPS equipment
BUS_UUID	Vehicle number	Unique identification of a vehicle
LINE_UUID	Line number	Unique identification of a bus line
LINE_NAME	Line name	Name of a bus line
DATA_TYPE	Type of data	Arriving/leaving a station, in-route
RECORD_TIME	GPS time	2018-03-01 13:17:54
LNG	Longitude	Longitude of a turning point
LAT	Latitude	Latitude of a turning point
ALT	Altitude	Altitude of a turning point
GPS_SPEED	GPS speed	GPS speed of the vehicle
ROTATE_ANGLE	Rotation angle	Rotation angle of a turning point

stations. Section 5 conducts a case study in Suzhou, China, to validate the performance of the proposed algorithm. Finally, we conclude the paper with some remarks and perspectives.

2. Data Description

Three crucial datasets can be obtained in the urban transit system, i.e., smart card data, bus GPS data, and bus station information data. Smart card data records the swiping time, vehicle number, and other transaction-related information of passengers, which contains mainly temporal information. In most cases, the station at which a passenger boards is not recorded. Bus GPS data records the trajectories of bus vehicles, which contain both spatial and temporal information. Bus station data records the static positions of bus stations of each bus line, which only contain the spatial information. The raw data of bus GPS and bus station location sometimes do not match because of different coordinate systems. In the rest of this section, the detailed description of these datasets is given.

2.1. Smart Card Data. The smart card data is recorded by the card swiping terminal equipped on each bus. When passengers board the vehicle, their trip information will be recorded. The recorded fields and corresponding description are listed in Table 1, including the record ID, card ID, card type, card swiping time, bus line number, and vehicle number, which are mainly temporal information.

2.2. Bus GPS Data. In practice, most buses in large- and medium-sized cities have been equipped with GPS terminals, which can accurately collect the real-time positioning information, along with the bus line number and operating direction, which contains both spatial and temporal information.

Table 3: Structure of bus station location data.

Field	Field name	Description
STOP_UUID	Number of bus line	Unique identification of a bus stop
STOP_NAME	Name of stops	Name of a bus stop
STOP_TYPE	Type of bus stop	Upward or downward
LNG	Longitude	Longitude of a bus station
LAT	Latitude	Latitude of a bus station

As shown in Table 2, the data fields include the record ID, vehicle number, bus type, bus line number, bus line name, operating status, timestamp, longitude, latitude, altitude, running speed, and running angle. The system error of the GPS data is about 10–30 meters. However, in practice, some of the GPS equipment may be obsolete and cannot be corrected in time. The systematic error would reach 50 meters.

2.3. Bus Station Location Data. The bus network information is stored in relational database in the form of the bus route information and the location information of bus stations (see Table 3). In the relational database, bus lines and bus stations are regarded as entities. A bus station belongs to at least one bus line, and a bus line contains at least one bus station. The major data fields include line number, line name, line direction, the stating and end location, and line type.

3. Passenger Boarding Station Extraction

The first step of bus OD matrix estimation is to obtain the passenger's boarding station, i.e., identifying the origin of each trip. As mentioned in the previous section, the smart card data only records the transaction time, while the spatial

information of the boarding station is absent. As for the GPS data, only the vehicle trajectories are recorded, while the spatial relationship between trajectories and the location of stations is also unclear. Hence, it is necessary to match the smart card data to the bus GPS data to extract the passenger's boarding station. In this section, we first use the clustering technique to obtain the bus arrival time, which can be further matched to the smart card's transaction time, and then extract the passenger's boarding station.

3.1. Obtaining Bus Arrival Time Based on Density Clustering Algorithm

3.1.1. Feasibility Analysis of Density Clustering Algorithm. Through the spatial matching of the bus station data and bus GPS data, the bus arrival time at each station can be obtained. However, it is difficult to directly create reference from the coordinates in bus trajectories to the location of bus stations, in that differences exist in the geographic coordinate systems, and bus GPS data can suffer from interference. These uncertainties and disturbances can be formulated as follows:

$$\begin{aligned} \mathrm{lng}_{\mathrm{LS}}^{(i,j)} &= \mathrm{lng}_{\mathrm{GPS}}^{(i,j,k)} + C_{\mathrm{lng}} + \delta_{\mathrm{lng}}^{(i,j,k)}, \\ \mathrm{lat}_{\mathrm{LS}}^{(i,j)} &= \mathrm{lat}_{\mathrm{GPS}}^{(i,j,k)} + C_{\mathrm{lat}} + \delta_{\mathrm{lat}}^{(i,j,k)}, \end{aligned} \tag{1}$$

where $\mathrm{lng}_{\mathrm{LS}}^{(i,j)}$ and $\mathrm{lat}_{\mathrm{LS}}^{(i,j)}$ represent the coordinates of bus station j on bus line i recorded in the bus station data; $\mathrm{lng}_{\mathrm{LS}}^{(i,j)}$ and $\mathrm{lat}_{\mathrm{LS}}^{(i,j)}$ represent the coordinates of the k-th sample of the bus station j on bus line i in the bus GPS data; C_{lng} and C_{lat} represent the conversion errors between the geographic coordinate systems adopted by the two different data sources; and $\delta_{\mathrm{lng}}^{(i,j,k)}$ and $\delta_{\mathrm{lat}}^{(i,j,k)}$ represent the positioning errors of the k-th sample of bus station j on bus line i in the bus GPS data. Due to the existence of these errors, it is essential to ensure the correctness of the spatial clustering and matching of data from two different sources.

The clustering algorithm is an unsupervised learning method which is capable of classifying the data into several groups by analyzing the similarity and mutuality between observed samples. The density-based clustering algorithm can further investigate the connectivity of the data from the perspective of the local density of samples. For the bus GPS data, a vehicle usually decelerates when approaching the station and then stops if this station is not skipped. This motion pattern leads to the spatial aggregation of trajectory points near bus stations, which is suitable for the use of density-based clustering algorithm. In the following section, the density-based spatial clustering of applications with noise (DBSCAN) algorithm is applied to obtain the bus station based on the GPS data [25], while the errors corresponding to the direct data fusion can therefore be avoided.

3.1.2. Obtaining the Bus Arrival Time Based on DBSCAN Algorithm. The DBSCAN algorithm can be first applied on the bus GPS data, whereby the location of bus stations in the bus GPS data can be identified. Then, the bus GPS data is further matched with the bus station data based on the connectivity between trajectory points to generate bus arrival timetable without being affected by the errors from different data sources. The detailed steps of the proposed method for bus timetable generation are as follows:

Step 1: Density-based clustering

The DBSCAN algorithm involves two steps, namely, searching for core samples and generating clusters. Before executing the algorithm, the bus station data is first matched with the bus GPS data. In the DBSCAN algorithm, two key parameters need to be defined: ε, distance, and MinPts, the minimum number of points. The ε is defined to measure the density-reachable range, within which the points are considered as the neighborhood of the same cluster. The MinPts limits the maximum number of points in the same cluster. Through DBSCAN, the samples are split into two categories, i.e., core samples and noncore samples. Instead of randomly selecting a core sample from the complete dataset, we only select the sample from the bus GPS data as a clustering seed. The cluster grows from the seed to its neighboring samples, and the seed is then labelled as the primary core sample. By repeating these steps, the clusters and their primary core samples can be generated.

Step 2: Searching for the bus stations

After finding all the clusters in the dataset, the bus station corresponding with the samples needs to be determined. Firstly, if the sample is directly density reachable from a primary core sample, it will be labelled as the core sample of the cluster that the primary core sample belongs to and connected with the same bus station of the primary core sample. Secondly, if a sample is density reachable from exactly one primary core sample, it will also be matched with the bus station of the primary core sample. Thirdly, if a sample is density reachable from multiple primary core samples, the core samples falling in its ε-neighborhood are placed in the candidate set. Then, we count the number of samples that are density reachable from each candidate core sample, and the one with most density-reachable samples will be matched with the selected sample. The bus station of that sample will be the same as the corresponding bus station of that core sample.

Step 3: Obtaining the bus arrival time

The primary core samples identified by the DBSCAN algorithm and the corresponding bus stations can be used to generate the bus timetable. The time recorded by the primary core samples is the arrival time at those bus stations. Based on the algorithm above, the matching of the bus GPS data and bus station data in the spatial dimension can be realized, and an accurate bus timetable, which is fundamental to bus OD estimation and records the exact time when each vehicle arrives at each bus station, can be generated.

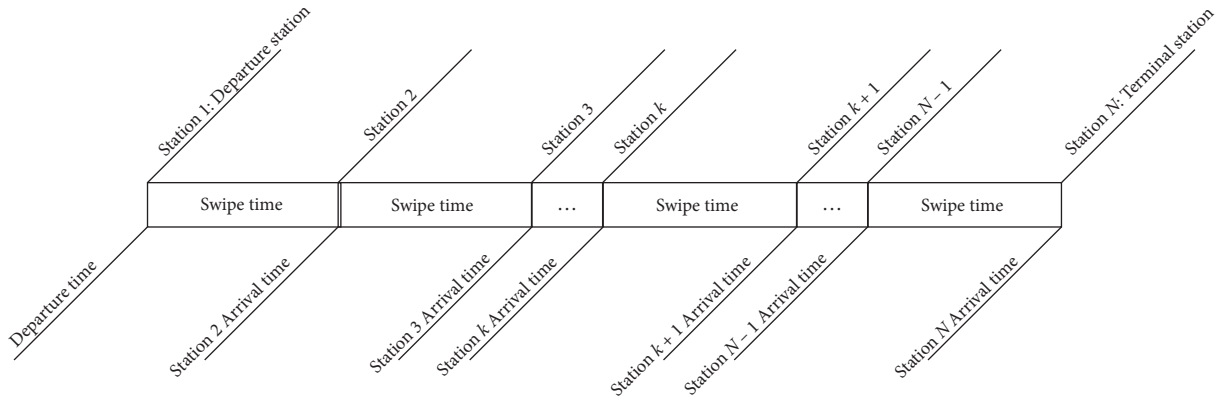

Figure 1: Relationship of card swiping time and bus arrival time.

3.2. Boarding Station Identification Based on Temporal Matching. Through the effective matching of the bus GPS data and the bus station data, the bus arrival time has been obtained. Subsequently, the boarding station can be identified by building connections between the smart card data and the bus arrival time in the temporal dimension.

3.2.1. Temporal Matching Algorithm. The premise of applying the above matching idea is to align the recorded time of the smart card data with the bus arrival time. Since the card swiping system and the GPS terminal installed on the bus operate independently, a fixed time difference between the recorded time values, referred to as the system time difference, might exist. Therefore, before identifying the boarding stations, the data from the two sources should be adjusted to ensure that they are consistent in the temporal dimension.

Generally, the time recorded in the smart card data is merely the transaction time rather than the true boarding time. As shown in Figure 1, the swiping behavior of bus passengers regularly occurs during the time the bus takes from the boarding station to the next station.

The system time difference between the smart card data and the bus GPS data will lead to the following two situations:

(1) The system time of the smart card data is earlier than that of the bus GPS data system. In such case, it is possible that the boarding station is mistakenly identified as the previous boarding station, resulting in a lower number of boardings at the current station and a higher number of boardings at the previous station.

(2) The system time of the smart card data system time is later than that of the bus GPS data system. In this case, it is possible that the boarding station is mistakenly identified as the next station, resulting in a higher number of boardings at the current station and a lower number of boardings at the previous station.

Assuming the system time difference between smart card system and bus GPS system is Δt, and this difference is constant for the two data sources. The system time difference Δt can be expressed as

$$\Delta t = \operatorname{sign}\left(\Delta t^{(i,d)}\right) \times \max_{i,d} \left|\Delta t^{(i,d)}\right|, \tag{2}$$

$$\Delta t^{(i,d)} = t_{\mathrm{IC}}^{(i,d)} - t_{\mathrm{GPS}}^{(i,d)}, \tag{3}$$

where $\Delta t^{(i,d)}$ and $\Delta t^{(j,d)}$ are subject to

$$\Delta t^{(i,d)} \times \Delta t^{(j,d)} > 0, \quad \forall \Delta t^{(i,d)}, \Delta t^{(j,d)} \in T, \tag{4}$$

$$\Delta t^{(i,d)} < \Delta T, \tag{5}$$

where $\Delta t^{(i,d)}$ represents the system time difference of i-th bus routes on day, $t_{\mathrm{IC}}^{(i,d)}$ represents the time of the first transaction of the i-th bus routes on day d, $t_{\mathrm{GPS}}^{(i,d)}$ represents the time of the first bus arrival of the i-th bus routes on day d, T represents the collection of $t_{\mathrm{IC}}^{(i,d)}$, ΔT represents the threshold of the maximum system time difference, and $\operatorname{sign}(\cdot)$ indicates whether the value is positive or negative.

Equation (2) takes the maximum time difference between the card swiping time of the first bus on that day and the first bus arrival time as the system time difference. Constraint (4) guarantees that all calculated time difference values take the same sign. Constraint (5) ensures that the first transaction recorded in the smart card data corresponds to the boarding at the first station of that bus route.

In practice, a minor lag might exist between the transaction time of the first passenger and the bus arrival time at the terminus. This can be addressed by simply adding a lag coefficient $\delta^{(i,d)}$ to equation (3), formulated as

$$\Delta t^{(i,d)} = t_{\mathrm{IC}}^{(i,d)} - t_{\mathrm{GPS}}^{(i,d)} - \delta^{(i,d)}, \tag{6}$$

where $\delta^{(i,d)}$ represents a lag coefficient satisfying a specific probability distribution, which can be calibrated according to the actual condition. Finally, using the system correction algorithm, the time of the smart card data can be revised as

$$t_{\mathrm{IC}}^{(i,d)} = t_{\mathrm{IC}}^{(i,d)} - \Delta t, \tag{7}$$

where $t_{\mathrm{IC}}^{(i,d)}$ represents the card swiping time on day d of bus line i after correction.

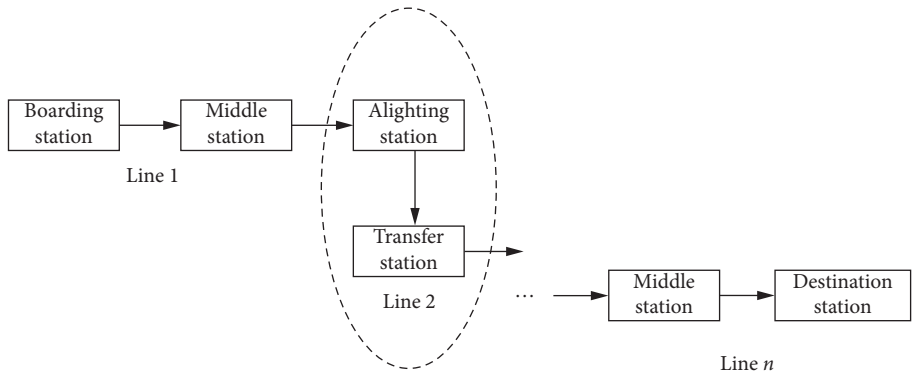

Figure 2: Typical bus trip chain.

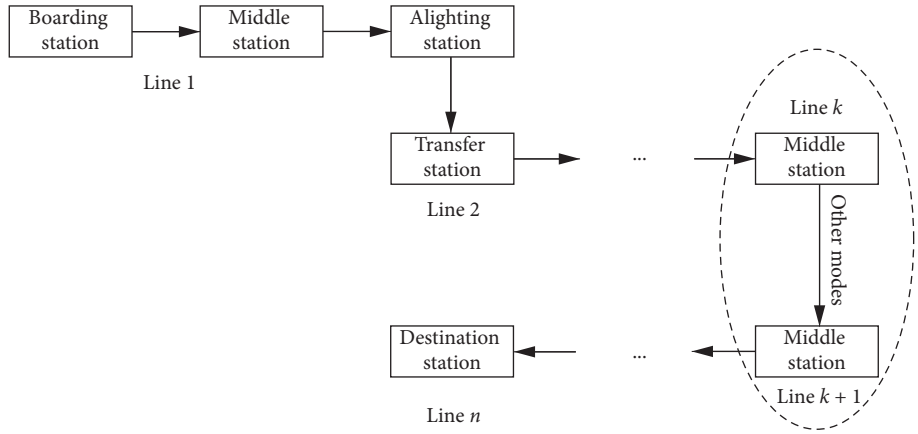

Figure 3: Unclosed bus trip chain.

3.2.2. Framework of Boarding Station Recognition. After correcting the system time difference based on the characteristics of multisource data, the overall framework of the boarding station recognition algorithm proposed in this paper can be summarized in two steps.

Step 1: System time correction

The transaction time of the first passenger recorded in the smart card data and the first bus arrival data each day for all bus lines are extracted for system time correction. According to the system time correction algorithm introduced in the previous section, the corrected smart card data can be obtained by Constraint (5).

Step 2: Boarding station identification

The boarding stations are identified based on the corrected smart card data and bus arrival timetable data according to the condition that the card swiping behavior happens in the time interval between two stations. In other words, the transaction time should be later than the bus arrival time at a station but earlier than the bus arrival time of the next station, where the first station here is regarded as the boarding station of that passenger.

4. Estimation Method of the Alighting Station Based on Trip Chaining

4.1. Definition of Bus Trip Chaining. Relevant research indicates that single trip of urban residents is the basic unit of the trip chaining. Bus trip chaining can be further defined as a process where residents take bus travel and form at least one spatial connection to neighboring travel. A bus trip chain can either be closed or unclosed. As shown in Figure 2, a typical and complete bus trip chain consists of the origin station, bus lines, middle stations, transfer stations, transfer bus lines, and the destination station. It has the following two features: (1) the passenger's alighting station of his/her non-last bus travel is spatially connected to the boarding station of his next bus travel on the same day and (2) the passenger's alighting station of his/her latest travel is spatially connected to the boarding station of his first travel.

However, in a multimodal transit system, passengers intend to use a combined travel pattern including different travel modes (e.g., bus, metro, taxi). Hence, the bus trip chain of a passenger is sometimes unclosed. As shown in Figure 3, the "alighting station" and "transfer station" in the cycle may not have clear spatial connection within the same

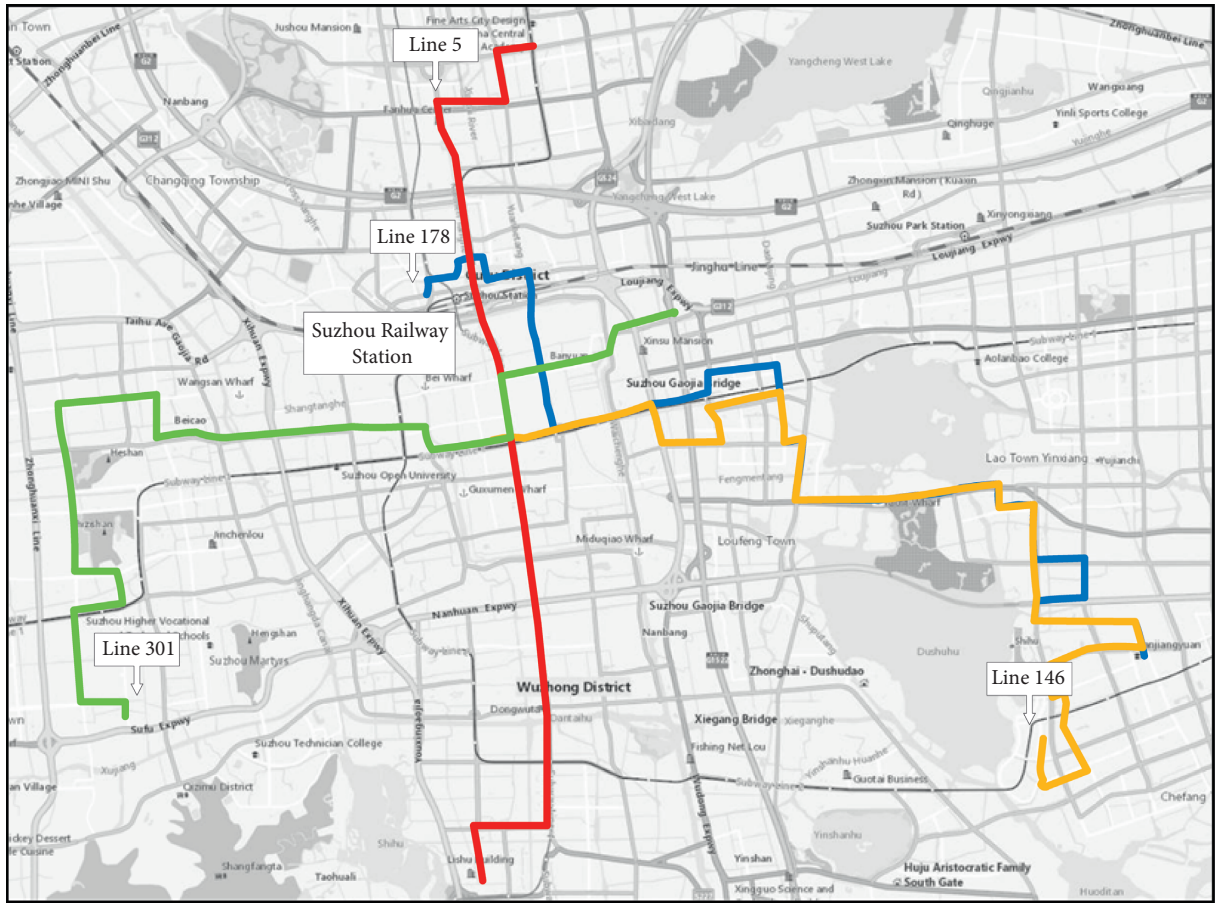

Figure 4: Layout of selected bus lines.

Table 4: Description of selected bus lines.

No. of bus line	Description	Direction
5	Main commuter line of downtown	S-N
301	Main commuter line of downtown	E-W
178	Connecting line between railway station and suburb	E-W
146	Main commuter line of suburb	S-N

day, which indicates that the passenger may choose other travel modes between these two stations. In this regard, the unclosed trip chain can be restored by the passenger's historical trip data to distinguish his/her regular or occasional trip.

In a bus trip chain, a typical passenger picks up the bus line L_1 at bus station S_1^1 and gets off at station S_1^2. Then, he/she transfers to bus line L_2 via station S_1^2 and gets off at station S_2^2. After n times of transfer, he/she arrives at the destination station S_{n+1}^2. The distance between the k-th alighting station S_k^2 and the $k+1$-th alighting station S_{k+1}^1 should be shorter than the acceptable walking distance. In summary, bus trip chaining should meet the following constraints: (1) component constraint: a bus trip chain must consist of at least two complete bus travels; (2) temporal constraint: the boarding time in last trip must be earlier than the boarding time in next trip; (3) spatial constraint: the distance between the alighting station in last trip and the boarding station in next trip should be shorter than the typically acceptable walking distance (e.g., 500 –1500 meters); and (4) transfer constraint: the transfer times in a bus trip chain should be lower than the acceptable frequency.

4.2. Alighting Stations Estimation Algorithm. On the one hand, it is difficult to estimate those unchained trips due to their irregularity and unpredictability. On the other hand, bus network planning usually focuses on the main travel behavior such as commuting. Therefore, estimating the alighting station for the chained trips can not only extract the spatiotemporal characteristics and OD matrix of most passenger's bus travel, but also meet the needs of bus network planning and design. By stripping and reforming the above dataset, the closed bus trip chain dataset and the unbroken part of unclosed bus trip chain dataset are integrated to form the dataset of chained trips. Meanwhile, the broken part of the unclosed bus trip chain data and the remaining data are added to the dataset of unchained trips.

4.2.1. Alighting Stations Estimation of Non-Last Bus Travel. According to the result of boarding station identification, it is assumed that a passenger picks up on the $j^{(L_i)}$ station of

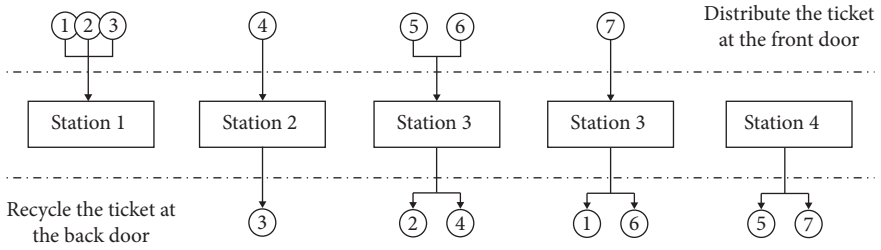

Figure 5: Illustration of the ticket recycling survey method for a single bus line.

(a) (b)

Figure 6: Distributing and recycling of the ticket at the front and back doors, respectively.

bus line L_i and his/her next boarding station is $j^{(L_i+1)}$ station of bus line L_{i+1}. Based on the spatial constraint and minimum station distance assumption, passenger would choose the alighting station which is the nearest to the next boarding station. Thus, the station which is the nearest to the $j^{(L_i+1)}$ station of line L_{i+1} from the stations in the downstream of the $j^{(L_i)}$ station online L_i is chosen as the alighting station of the last travel.

To ensure the effectiveness of data processing, the selection of the stations on bus line L_i where its distance from the $j^{(L_i+1)}$ station on bus line L_i is shorter than the acceptable walking distance, is recommended. A table storing such transfer topology information is calculated in advance for subsequent selection and processing.

4.2.2. Alighting Stations Estimation of Last Bus Travel. According to the result of boarding station identification, it is assumed that a passenger picks up on the $j^{(L_n)}$ station of bus line L_n in his/her last travel of the day, and his/her boarding station in first travel of the day is $j^{(L_1)}$ station of bus line L_1. Based on the minimum station distance assumption and commuting rules, passenger would choose the alighting station which is the nearest to the first boarding station on that day. Thus, the station which is the nearest to the $j^{(L_1)}$ station of line L_1 from the stations in the downstream of the $k^{(L_n)}$ station is chosen as the alighting station of the last travel of the day.

Table 5: Statistical description of the MAE.

Statistics	Boarding station identification Value	Alighting station identification Value
Average	1.013	1.206
Standard	0.443	0.496
Minimum	0.140	0.264
25th percentile	0.669	0.834
Median	1.007	1.144
75th percentile	1.265	1.502
Maximum	2.256	2.550

4.2.3. Alighting Stations Estimation of Unclosed Bus Trip Chain. Unclosed bus trip chain can be split into the broken and unbroken parts. If the broken part occurs in the middle of the chain, the broken parts are usually recovered by the historical data while the remaining unbroken part can be regarded as a continuous bus trip chain. If the broken chain is in the head and tail of the chain, the alighting stations of non-last bus travel can be estimated in the same way as that used in a closed bus trip chain. Instead of the recovery method based on historical data, another method named chaining extension which tries to establish a spatial connection between the neighboring bus travels in neighboring days is preferred due to its higher confidence.

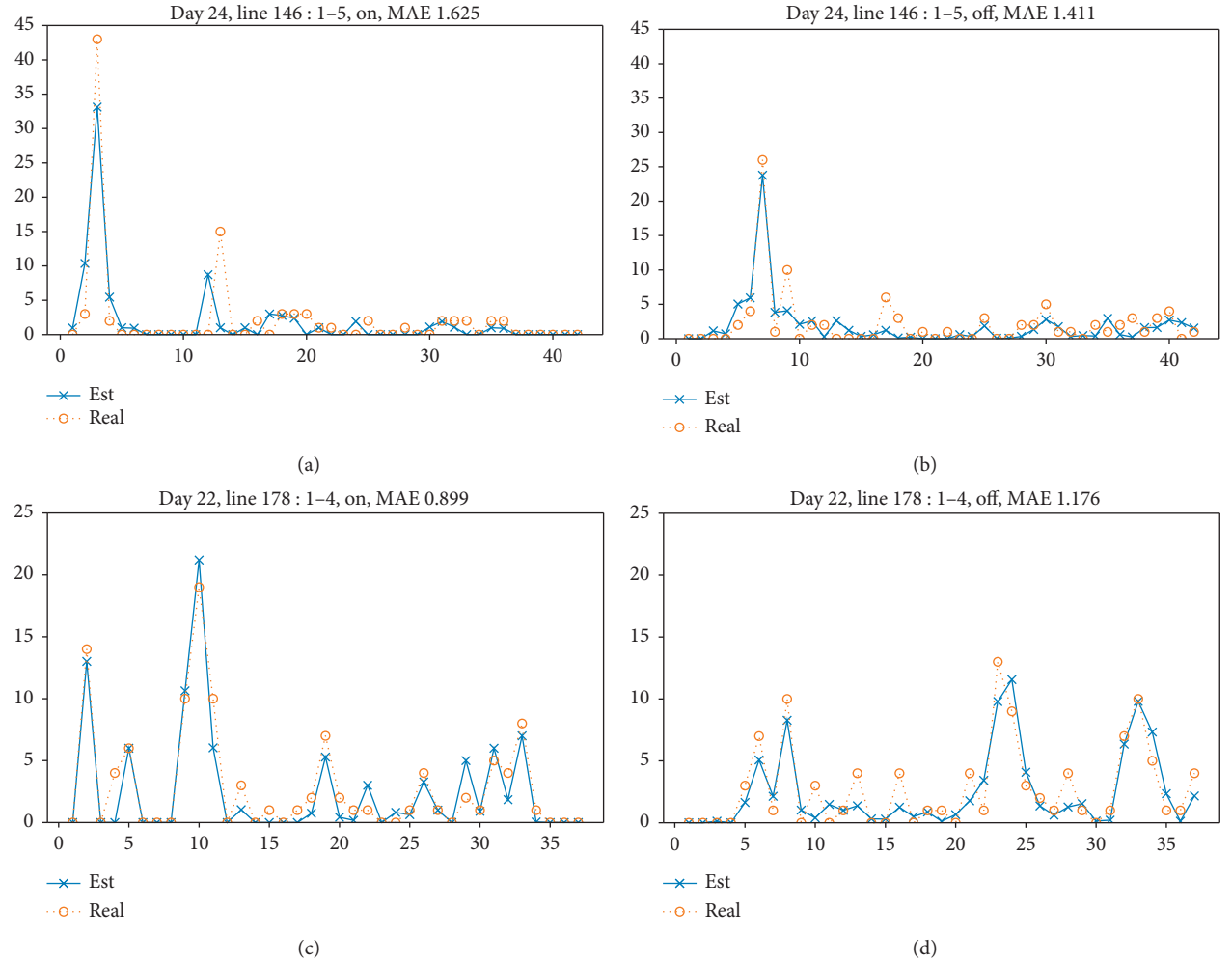

Figure 7: True values and estimated values of boarding and alighting numbers. (a) Line146: boarding passengers. (b) Line146: alighting passengers. (c) Line178: boarding passengers. (d) Line178: alighting passengers.

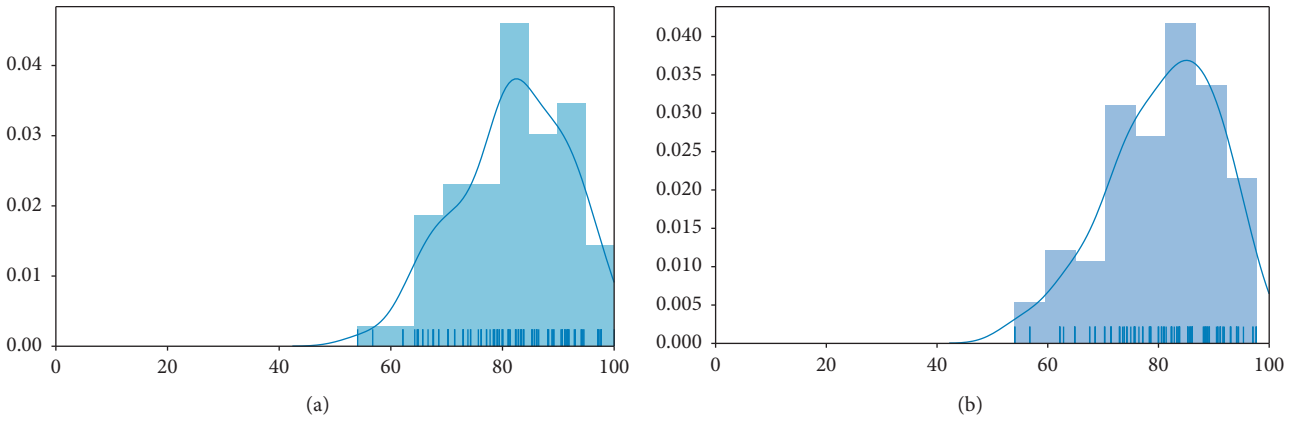

Figure 8: Percentage of stations where numbers of boarding and alighting passengers are accurately calculated. (a) Boarding station identification. (b) Alighting station estimation.

After reasonably dividing the bus datasets that have been effectively identified on the boarding station, we use the proposed algorithm based on the bus trip chain to estimate the alighting station and obtain the OD matrix.

5. Experiment and Verification

In order to verify the effectiveness of the proposed methodology, a survey was conducted in Suzhou, China, to collect the real bus travel OD information. As shown in Figure 4 four typical bus lines are selected, which have different functions (see Table 4). The ticket recycling survey method is used to collect the real number of passengers from station to station.

5.1. Ticket Recycling Survey Method. The ticket recycling survey method is a classical flow-up survey which has been widely applied to obtain the real passenger distribution along a bus route [40]. As shown in Figures 5 and 6, each passenger would receive a ticket that records the number of the station where s/he boards the bus at the front door. This ticket will be collected before the passenger alight from the back door. Though this method is time consuming and costly, it has a 100% sample rate which could obtain the real station-to-station trip information. The survey contains 72 round trips of these four lines. The total number of valid tickets is 10,551.

5.2. Accuracy Evaluation. By counting the number of board and alight passengers surveyed, the mean absolute error (MAE)of each class (i.e., from the origin station to the destination station) can be calculated. As a result, the MAE on each class is 1.326 passengers/station for numbers of boarding passengers and 1.299 passengers/station for numbers of alighting passengers. The detained statistical description of the MAE is presented in Table 5.

Figure 7 shows the verification results of bus line 146 (departure time: 8:25 am, October 24) and bus line 178 (departure time: 7:45 am, October 22). In order to find out whether the algorithm is accurate for most stations, this paper also proposes another checking indicator by calculating the percentage of the number of stations whose error is less than a certain threshold (number of passengers) to the total number of stations. It is assumed that the result is considered to be accurate when the estimated value differs from the actual value by two or less. As shown in Figure 8, the average percentage of the number of boarding stations is accurately identified as 81.8%, and this indicator is 80.9% for the alighting station estimation.

Furthermore, consider that there is a certain systematic error between the sampling time of the card reader and the POS time which is difficult to obtain. Table 6 lists the accurate inference percentage of each bus line and adds the limited dynamic time warping (LDTW). LDTW is usually used to evaluate time series similarity. In this study, the constraint of deformation distance is added on the standard dynamic time deformation. This means the value a_t in the time series can only match the adjacent time interval a_{t-1} and a_{t+1}. This indicator has better robustness than an absolute error by considering the relevance of the attraction adjacent stations to passenger flow.

Table 6: Estimation accuracy for boarding and departing passengers.

Line	Boarding station identification		Alighting station estimation	
	Accuracy (%)	LDTW	Accuracy (%)	LDTW
5	82.0	9.171	82.2	9.508
146	85.3	9.815	86.7	8.916
178	85.4	12.475	81.6	12.150
301	71.7	14.031	70.0	14.898

Another practical significance of using LDTW for evaluation is that it can prove the prediction accuracy of the algorithm for the overall trend. LDTW can be more accurate to estimate a small area instead of one station due to the fact that the adjacent stations are usually close enough to be considered as a small district. The verification results show that the accuracy of line 5 and line 146 is above 80%, and the corresponding LDTW is less than 10. The performance on line 301 is comparatively worse, but the accurate percentage is still above 70%, and the LDTW is less than 15. Tables 7 and 8 present a detailed comparison between the estimated and observed results for 10 stations with the highest passenger volume. It shows that the estimation values of the boarding passengers are almost less than 10 passengers, while the alighting passengers are less than 15 passengers.

5.3. Accuracy Evaluation of OD Matrix Estimation. By combining the results of boarding and alighting station estimation, the accuracy of OD matrix in each bus class can be calculated (the statistics do not include the OD with estimated and actual value being 0). The mean absolute error of the OD estimation is 0.581. Figure 9 shows the distribution of OD estimation accuracy. The accuracy threshold of Figure 9(a) is 2 passengers difference between the inferred value of the OD and the actual value, and the average accuracy is 94.3%. Figure 9(b) shows the accuracy threshold to 1 person, and the average accuracy is 72.8%. Table 9 lists the MAE of OD estimation in different bus lines. It can be seen from the table that the accurate percentage of each line (the error is less than 2 passengers) is above 90%, while the mean absolute error (MAE) is basically below 0.7. By comparing the estimated value with the real data of the bus travel survey, it can be found that the proposed algorithm performs well both in the identification of boarding and alighting station and in the overall OD matrix estimation. It can effectively obtain the passenger flow of each station and OD of different bus lines.

Table 7: Comparison between the estimated and observed values of the boarding passengers.

Date	No. of line	No. of bus	No. of station	Est. value	Obs. value	Error
24/Oct	146	5	3	33.125	43	9.875
23/Oct	178	2	2	28.817	39	10.183
24/Oct	178	3	2	44.720	38	6.720
23/Oct	146	1	14	32.657	34	1.343
25/Oct	5	9	4	22.418	33	10.582
23/Oct	178	3	2	35.516	32	3.516
22/Oct	146	2	14	24.355	30	5.645
22/Oct	146	0	14	19.633	24	4.367
24/Oct	146	1	14	19.281	24	4.719
23/Oct	178	2	5	18.448	23	4.552

Table 8: Comparison between the estimated and observed values of the alighting passengers.

Date	No. of line	No. of bus	No. of station	Est. value	Obs. value	Error
24/Oct	146	1	16	31.341	36	4.659
23/Oct	178	2	6	19.012	29	9.988
24/Oct	146	5	7	23.772	26	2.228
23/Oct	178	3	8	11.125	26	14.875
25/Oct	146	1	16	19.298	24	4.702
23/Oct	146	0	16	8.140	24	15.860
22/Oct	178	4	6	15.254	24	8.746
22/Oct	5	9	9	10.822	24	13.178
24/Oct	301	0	36	7.352	23	15.648
23/Oct	5	2	29	11.432	23	11.568

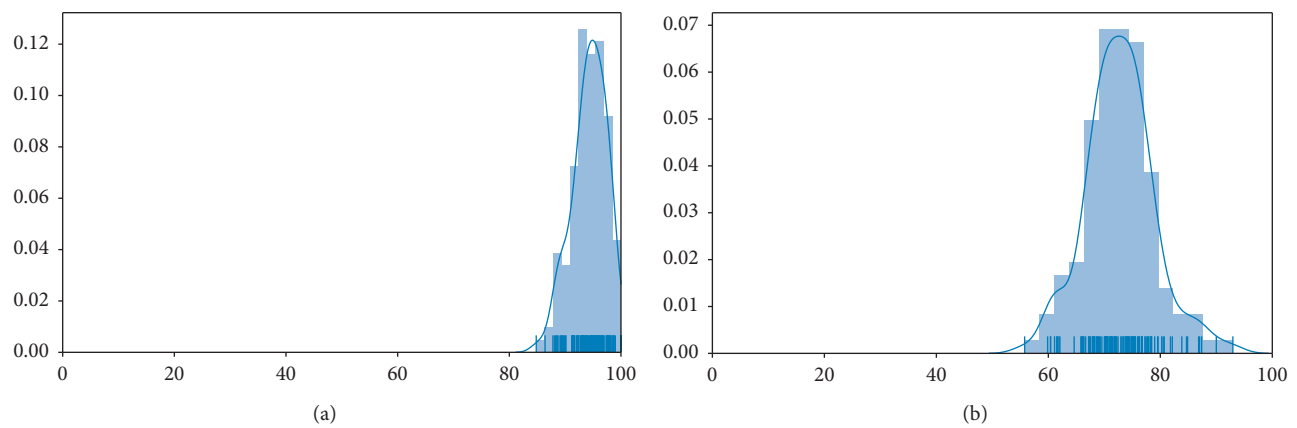

Figure 9: Percentage of accurate OD estimation. (a) Error less than 2 passengers. (b) Error less than 1 passenger.

Table 9: Descriptive statistics of OD matrix.

No. of line	Rate of accurate OD (error < 2) (%)	MAE
5	94.9	0.543
146	95.2	0.554
178	94.3	0.626
301	92.6	0.630

6. Conclusions

The accurate estimation of bus OD matrix is essential for the planning and management of urban bus system. This paper proposes a framework for estimating the OD matrix, including a boarding station identification algorithm and an alighting station inference algorithm. The boarding station identification algorithm allows for the mismatch of system time between the smart card system and the bus GPS system. These two datasets are aligned in the temporal dimension through the density-based clustering algorithm. Then, based on the identified chained and unchained trips, an alighting station inference algorithm based on trip chaining is designed. Above 80% of alighting stations could be identified by using the proposed estimation algorithm, which increases the identification rate by 10% compared with previous studies [33, 36, 38]. Finally, the accuracy of the proposed algorithm is evaluated on the data collected by the field survey in Suzhou, China. The results show that the proposed methodology could obtain an accuracy level above 90%.

Also, this algorithm can be further improved in the following aspects. First, the proposed framework works only

within the bus network. However, with the development of multimodal urban transport system, travelers' trip chains can include other travel modes such as light rail, subways, and even shared bicycles. Second, more data sources can be introduced to compensate for the incomplete market share of smart cards.

Data Availability

The data used to support the findings of this study have not been made available because of confidential issues.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This study was supported by the National Key Research and Development Program of China (No. 2018YFB1600900), the Key Project (No. 51638004) of the National Natural Science Foundation of China, the Scientific Research Foundation of Graduate School of Southeast University (No. YBPY1835), and DiDi Gaia Research Collaboration Initiative. The authors would like to thank Cheng Lyu and Yunyang Shi for their efforts in data preprocessing.

References

[1] Y. Liu, Z. Liu, and R. Jia, "DeepPF: a deep learning based architecture for metro passenger flow prediction," *Transportation Research Part C: Emerging Technologies*, vol. 101, pp. 18–34, 2019.

[2] E. Chen, Z. Ye, C. Wang, and M. Xu, "Subway passenger flow prediction for special events using smart card sata," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–12, 2019.

[3] D. Huang, Z. Liu, X. Fu, and P. T. Blythe, "Multimodal transit network design in a hub-and-spoke network framework," *Transportmetrica A: Transport Science*, vol. 14, no. 8, pp. 706–735, 2018.

[4] D. Huang, Z. Liu, P. Liu, and J. Chen, "Optimal transit fare and service frequency of a nonlinear origin-destination based fare structure," *Transportation Research Part E: Logistics and Transportation Review*, vol. 96, pp. 1–19, 2016.

[5] Z. Liu, S. Wang, B. Zhou, and Q. Cheng, "Robust optimization of distance-based tolls in a network considering stochastic day to day dynamics," *Transportation Research Part C: Emerging Technologies*, vol. 79, pp. 58–72, 2017.

[6] Y. Bie, X. Xiong, Y. Yan, and X. Qu, "Dynamic headway control for high-frequency bus line based on speed guidance and intersection signal adjustment," *Computer-Aided Civil and Infrastructure Engineering*, vol. 35, no. 1, pp. 4–25, 2020.

[7] X. Fu and W. H. K. Lam, "Modelling joint activity-travel pattern scheduling problem in multi-modal transit networks," *Transportation*, vol. 45, no. 1, pp. 23–49, 2018.

[8] C. Wang, Z. Ye, E. Chen, M. Xu, and W. Wang, "Diffusion approximation for exploring the correlation between failure rate and bus-stop operation," *Transportmetrica A: Transport Science*, vol. 15, no. 2, pp. 1306–1320, 2019.

[9] Y. Pan, S. Chen, F. Qiao, S. V. Ukkusuri, and K. Tang, "Estimation of real-driving emissions for buses fueled with liquefied natural gas based on gradient boosted regression trees," *Science of the Total Environment*, vol. 660, pp. 741–750, 2019.

[10] S. Tao, D. Rohde, and J. Corcoran, "Examining the spatial-temporal dynamics of bus passenger travel behaviour using smart card data and the flow-comap," *Journal of Transport Geography*, vol. 41, pp. 21–36, 2014.

[11] Y. Ji, Y. Fan, A. Ermagun, X. Cao, W. Wang, and K. Das, "Public bicycle as a feeder mode to rail transit in China: the role of gender, age, income, trip purpose, and bicycle theft experience," *International Journal of Sustainable Transportation*, vol. 11, no. 4, pp. 308–317, 2017.

[12] M. Du and L. Cheng, "Better understanding the characteristics and influential factors of different travel patterns in free-floating bike sharing: evidence from Nanjing, China," *Sustainability*, vol. 10, no. 4, p. 1244, 2018.

[13] J. J. Barry, R. Newhouser, A. Rahbee, and S. Sayeda, "Origin and destination estimation in New York City with automated fare system data," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1817, no. 1, pp. 183–187, 2002.

[14] Y. Yuan, M. Yang, J. Wu, S. Rasouli, and D. Lei, "Assessing bus transit service from the perspective of elderly passengers in Harbin, China," *International Journal of Sustainable Transportation*, vol. 13, no. 10, pp. 761–776, 2019.

[15] M. Utsunomiya, J. Attanucci, and N. Wilson, "Potential uses of transit smart card registration and transaction data to improve transit planning," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1971, no. 1, pp. 118–126, 2006.

[16] L. Sun, A. Tirachini, K. W. Axhausen, A. Erath, and D.-H. Lee, "Models of bus boarding and alighting dynamics," *Transportation Research Part A: Policy and Practice*, vol. 69, pp. 447–460, 2014.

[17] Y. Ji, X. Ma, M. Yang, Y. Jin, and L. Gao, "Exploring spatially varying influences on metro-bikeshare transfer: a geographically weighted Poisson regression approach," *Sustainability*, vol. 10, no. 5, p. 1526, 2018.

[18] T. Spurr, R. Chapleau, and D. Piché, "Use of subway smart card transactions for the discovery and partial correction of travel survey bias," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2405, no. 1, pp. 57–67, 2014.

[19] M.-P. Pelletier, M. Trépanier, and C. Morency, "Smart card data use in public transit: a literature review," *Transportation Research Part C: Emerging Technologies*, vol. 19, no. 4, pp. 557–568, 2011.

[20] A. Kiyohiro, K. Yamaguchi, H. Gao, H. Nakamura, and T. Mine, "Customer behavior analysis on after getting off the train based on usage histories of smart IC card," in *Proceedings of the 3rd International Conference on Advanced Applied Informatics*, Kitakyushu, Japan, August 2014.

[21] Q. Ouyang, Y. Lv, Y. Ren, J. Ma, and J. Li, "Passenger travel regularity analysis based on a large scale smart card data," *Journal of Advanced Transportation*, vol. 2018, Article ID 9457486, 11 pages, 2018.

[22] N. Ebadi, J. E. Kang, and S. Hasan, "Constructing activity-mobility trajectories of college students based on smart card transaction data," *International Journal of Transportation Science and Technology*, vol. 6, no. 4, pp. 316–329, 2017.

[23] Y. Ji, J. Zhao, Z. Zhang, and Y. Du, "Estimating bus loads and OD flows using location-stamped farebox and Wi-Fi signal

data," *Journal of Advanced Transportation*, vol. 2017, Article ID 6374858, 10 pages, 2017.

[24] X.-L. Ma, Y.-H. Wang, F. Chen, and J.-F. Liu, "Transit smart card data mining for passenger origin information extraction," *Journal of Zhejiang University Science C*, vol. 13, no. 10, pp. 750–760, 2012.

[25] X. Ma, Y.-J. Wu, Y. Wang, F. Chen, and J. Liu, "Mining smart card data for transit riders' travel patterns," *Transportation Research Part C: Emerging Technologies*, vol. 36, pp. 1–12, 2013.

[26] T. Kusakabe and Y. Asakura, "Behavioural data mining of transit smart card data: a data fusion approach," *Transportation Research Part C: Emerging Technologies*, vol. 46, pp. 179–191, 2014.

[27] D. Chen, "Research on traffic flow prediction in the big data environment based on the improved RBF neural network," *IEEE Transactions on Industrial Informatics*, vol. 13, no. 4, pp. 2000–2008, 2017.

[28] J. Bao, P. Liu, X. Qin, and H. Zhou, "Understanding the effects of trip patterns on spatially aggregated crashes with large-scale taxi GPS data," *Accident Analysis & Prevention*, vol. 120, pp. 281–294, 2018.

[29] L. Li, J. Zhang, Y. Wang, and B. Ran, "Missing value imputation for traffic-related time series data based on a multi-view learning method," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 8, pp. 2933–2943, 2019.

[30] A. A. Nunes, T. G. Dias, and J. F. e Cunha, "Passenger journey destination estimation from automated fare collection system data using spatial validation," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 1, pp. 133–142, 2016.

[31] D. Zhang, X. Zhang, and J. Wang, "Commuter travel identification based on bus IC data," *Procedia-Social and Behavioral Sciences*, vol. 96, pp. 1547–1555, 2013.

[32] J. M. Farzin, "Constructing an automated bus origin-destination matrix using farecard and global positioning system data in São Paulo, Brazil," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2072, no. 1, pp. 30–37, 2008.

[33] M. Trépanier, N. Tranchant, and R. Chapleau, "Individual trip destination estimation in a transit smart card automated fare collection system," *Journal of Intelligent Transportation Systems*, vol. 11, no. 1, pp. 1–14, 2007.

[34] N. Nassir, A. Khani, S. G. Lee, H. Noh, and M. Hickman, "Transit stop-level origin-destination estimation through use of transit schedule and automated data collection system," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2263, no. 1, pp. 140–150, 2011.

[35] W. Wang, J. Attanucci, and N. Wilson, "Bus passenger origin-destination estimation and related analyses using automated data collection systems," *Journal of Public Transportation*, vol. 14, no. 4, pp. 131–150, 2011.

[36] J. Zhao, A. Rahbee, and N. H. M. Wilson, "Estimating a rail passenger trip origin-destination matrix using automatic data collection systems," *Computer-Aided Civil and Infrastructure Engineering*, vol. 22, no. 5, pp. 376–387, 2007.

[37] C. Seaborn, J. Attanucci, and N. H. M. Wilson, "Analyzing multimodal public transport journeys in London with smart card fare payment data," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2121, no. 1, pp. 55–62, 2009.

[38] A. Cui, *Bus Passenger Origin-Destination Matrix Estimation Using Automated Data Collection Systems*, Massachusetts Institute of Technology, Cambridge, MA, USA, 2006.

[39] K. Lu, A. Khani, and B. Han, "A trip purpose-based data-driven alighting station choice model using transit smart sard data," *Complexity*, vol. 2018, Article ID 3412070, 14 pages, 2018.

[40] M. An, X. Chen, and Z. Li, "Transit OD matrix estimation based on trickey survey method," *Journal of Transportation System Engineering and Information Technology*, vol. 10, no. 1, pp. 170–176, 2009.