Hindawi
Mathematical Problems in Engineering
Volume 2021, Article ID 6464980, 15 pages
https://doi.org/10.1155/2021/6464980

Research Article

Alighting Stop Determination of Unlinked Trips Based on a Two-Layer Stacking Framework

Ziwei Cui[1,2] **Cheng Wang**[1] **Yueer Gao**[3] **Dingkang Yang**[1] **Wei Wei**[4] **Jianwei Chen,**[5] **and Ting He**[1]

[1]*College of Computer Science and Technology, Huaqiao University, Xiamen 361021, China*
[2]*School of Intelligent Systems Engineering, Sun Yat-sen University, Guangzhou 510006, China*
[3]*School of Architecture, Huaqiao University, Xiamen 361021, China*
[4]*School of Computer Science and Engineering, Xi'an University of Technology, Xi'an 710048, China*
[5]*Department of Mathematics and Statistics, San Diego State University, San Diego 92182, CA, USA*

Correspondence should be addressed to Cheng Wang; wangcheng@hqu.edu.cn

Received 21 August 2020; Revised 9 February 2021; Accepted 3 March 2021; Published 19 March 2021

Academic Editor: Haoran Zhang

Smart card data of conventional bus passengers are important basic data for many studies such as bus network optimization. As only boarding information is recorded in most cities, alighting stops need to be identified. The classical trip chain method can only detect destinations of passengers who have trip cycles. However, the rest of unlinked trips without destinations are hard to analyze. To improve the accuracy of existing methods for determining alighting stops of unlinked trips, a two-layer stacking-framework-based method is proposed in this work. In the first layer, five methods are used, i.e., high-frequency stop method, stop attraction method, transfer convenience method, land-use type attraction method, and improved group historical set method (I-GHSM). Among them, the last one is presented here to cluster records with similar behavior patterns into a group more accurately. In the second layer, the logistic regression model is selected to get the appropriate weight of each method in the former layer for different datasets, which brings the generalization ability. Taking data from Xiamen BRT Line Kuai 1 as an example, I-GHSM given in the first layer has proved to be necessary and effective. Besides, the two-layer stacking-framework-based method can detect all destinations of unlinked trips with an accuracy of 51.88%, and this accuracy is higher than that of comparison methods, i.e., the two-step algorithms with KNN (k-nearest neighbor), Decision Tree or Random Forest, and a step-by-step method. Results indicate that the framework-based method presented has high accuracy in identifying all alighting stops of unlinked trips.

1. Introduction

In the smart card system, the smart card allows the continuous collection of individualized transactional data about the use of public transport networks, and each card is always corresponding to one user. Therefore, smart card data have become a valuable source of information due to their larger scale than surveys permit and over long periods [1–3].

Smart card data can be used for data mining for identification of trip purpose [4], development of origin-destination matrices [5], estimation of vehicle load profile, and other network performance measures [6, 7]. All of these studies require known trip destinations [4–7]. However, in most cities' smart card system, only passengers' boarding transactions are recorded because users do not validate when they leave the buses. Thus, the alighting stops must be estimated [8, 9].

The trip chain method, described in the next section, is normally used to infer alighting stops [5]. Relying on the chaining of trips during the day, it is assumed that users will alight at the stop near the next boarding stop, and for the last trip of the day, the first boarding point of this day or the next day can be regarded as the next boarding stop to form the trip chain. But destinations that are not linked to the sequence cannot be estimated with this approach. These unlinked trips do not satisfy the approach criteria or they are alone within the day, which are the focus of this paper.

To identify the alighting stops of unlinked trips, most existing researchers use only one method or machine learning algorithm, and some other researchers use several fixed methods in sequence. However, which method or combination is the best for a different dataset is difficult to determine. This work proposes a two-layer stacking framework, which can get the appropriate contributions of several methods in a different dataset simultaneously, so this method has strong generalization ability and can detect all destinations with high accuracy.

Among existing methods of alighting stop determination of unlinked trips, the individual historical set method is widely used. In this method, an individual's records with alighting stops detected by the trip chain method are picked as the individual historical set, and then the stop with the highest alighting frequency in the historical dataset is the destination. But the individual historical set is small when passengers have fewer trips or more unlinked trips, which leads to limited use and a low identification rate [10, 11]. To expand the historical dataset, our previous study reported a method for alighting stop determination of transit passengers based on expanded history trip records [12], which is a step-by-step method. Firstly, the trip chain method with multisource data is presented, and data with destinations are used as the historical dataset. Then, the individual historical set method is used. Finally, for the rest data without alighting points, the group historical set method (GHSM) is used. Records of other passengers boarding at the same stop on the same line are selected as the group history dataset, and alighting stops are determined based on similar trips in the group history dataset. However, in the GHSM, the records' group can be clustered more carefully based on more features, so the improved group historical set method (I-GHSM) is proposed in this paper. Besides, different from researches dividing similar passengers into groups and making records of them in the same group [8, 13–17], this study takes each record as the smallest research unit and clusters them to make records with similar behavior patterns in the same group more directly and accurately.

The rest of the paper is organized as follows. Section 2 illustrates problem descriptions, data used, and related works. Methods to determine alighting stops of unlinked trips are introduced in Section 3. Section 4 manifests a case based on the presented method, and Section 5 holds the conclusions of this research.

2. Background

2.1. Problem Description. There must be an alighting stop for a user on the bus, but the alighting transaction is not recorded in most cities, so which stop is the alighting point is an important problem. The trip chain method is normally used to infer destinations [5], and it is introduced first as follows. However, some records' destinations cannot be identified by this method, and their alighting points are more difficult to determine. These records of unlinked trips are research objectives and their alighting point determinations are the study problems in this work.

The trip chain method is a classical model to estimate alighting points when a card with multiple records one day, and there are two hypotheses of this method [10, 11]:

(1) For a trip, users alight at the stop where the distance between the current alighting destination and the next boarding point is minimal, and the distance must be lower than a specified maximum value

(2) At the last trip of the day, users return to the first boarding station of this day or the next day, which is assumed to be the stop closest to home

Many algorithms are improved and generated from the trip chain method. For instance, Kumar et al. described a method to relax assumptions on various parameters, such as transferring walking distance threshold, buffering distance for selecting the boarding location, and developing a time window for selecting the vehicle trips [18]. Nassir et al. detected short activity locations among the stops visited on a trip chain and proposed a new heuristic to estimate the stop-level origins and destinations based on the traveler activities determination in the observed transactions [19]. Nunes et al. presented some new spatial validation features to increase destination inference results [20]. Compared with earlier methods on a heuristic basis, Sánchez-Martínez proposed a dynamic programming model to infer destinations with a generalized disutility minimization objective, which took the disutility of waiting, transferring, riding, and walking into consideration [21]. In terms of the use of data, the common trip chain method found passengers' trip chains based on the data of conventional bus IC cards, bus GPS, and static line station information. Cui et al. not only used the IC data of the conventional bus but also considered the IC card data of other public transport modes (such as the rail transit and the Bus Rapid Transit System (BRT)), to make the public transport trip chains more complete. The study case in Cui's research showed that using multisource IC card data can estimate more alighting stops with higher accuracy than single-source IC card data [12].

However, some passengers don't take public transportations frequently, and they have only one trip in a day or broken trips, which do not satisfy the approach criteria. Therefore, the destinations of unlinked trips cannot be estimated through the methods mentioned above, and they are the focus of this paper. The relationship of smart card data in the method based on a stacking framework can be found in Figure 1.

There are some examples of alighting stop determination in Figure 2:

(i) For the first trip, when a passenger boards at the k_1 stop of line A in the up direction, we can get the potential alighting stop $\{k_2, k_3, k_4, k_5, k_6\}$, and the k_5 stop will be the destination based on the 2nd boarding stop. Because the distance and time between two stops match rules in the trip chain method, these two trips build a trip chain.

(ii) For the second trip, we cannot confirm which potential alighting stop is the destination. Because

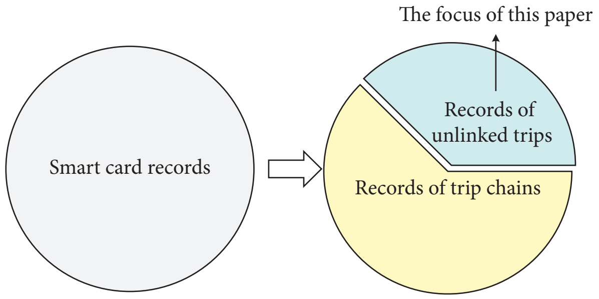

Figure 1: The relationship of smart card records in this paper.

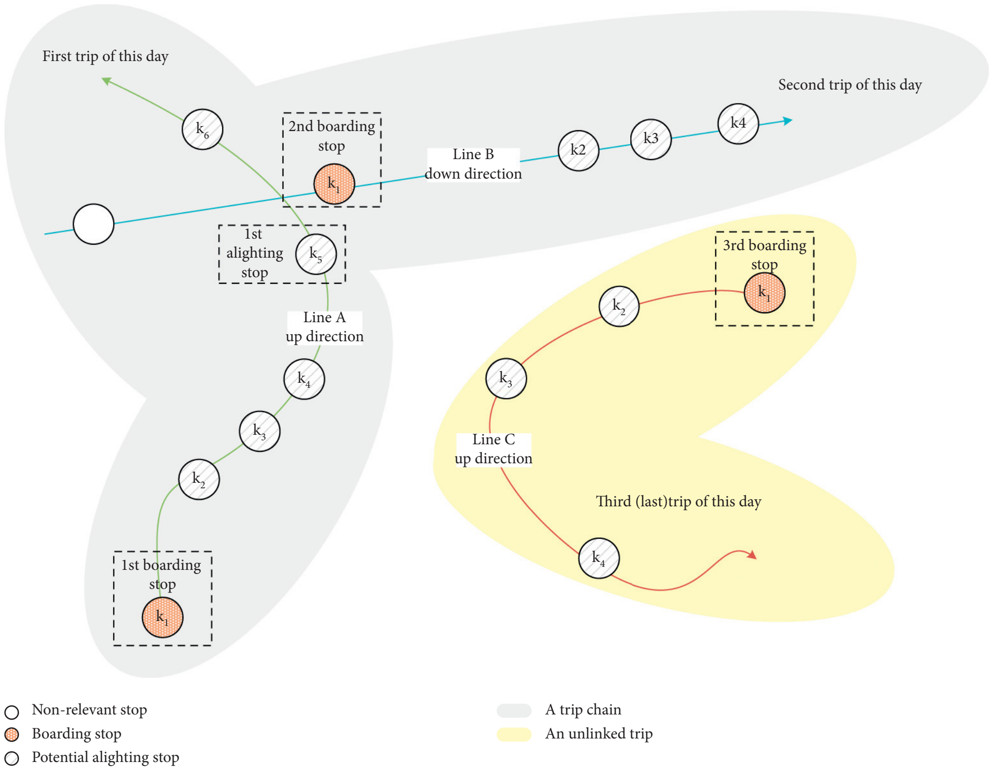

Figure 2: Some examples of alighting stop determination.

the next trip's line C is far away from line B, and the distances between the 3rd boarding stop and potential alighting stops in the second trip are not within the range, the second trip does not satisfy the approach criteria, and it is unlinked.

(iii) For the third (last) trip, similarly, we cannot detect the destination because the first boarding stop of the day is far away from potential alighting stops, and it is also an unlinked trip.

In this paper, we make contributions to identify alighting points of the second and third trips.

2.2. Data Description

2.2.1. Static Bus Stop Information. Static bus stop information contains the line number, the line's direction (up or down), and the index, name, longitude, and latitude of each stop along this direction of this line in Table 1. It is needed to count the number of lines passing through each stop, which is useful in the two-layer stacking framework method.

2.2.2. Bus GPS Data. As can be seen from Table 2, GPS data embedded on all buses have ten fields, line number, direction, stop index, stop name, bus license plate number, longitude, latitude, date, vehicle operation shift, and timestamp. For a smart card record, when the interval of the record's transaction time and the GPS timestamp is minimal, the corresponding stop index, name, longitude, and latitude will be regarded as the boarding stop information.

2.2.3. Land-Use Type Attraction Coefficient. The urban construction land with the bus stop as the center and within

Table 1: Static bus stop information.

Line no.	Direction	Stop index	Stop name	Longitude	Latitude
9 * 5	Up	1	A	118.056	24.6186
9 * 5	Up	2	B	118.058	24.6145
9 * 5	Up	3	C	118.060	24.6129

Table 2: Bus GPS data.

Line no.	Direction	Stop index	Stop name	Bus license plate no.	Longitude	Latitude	Date	Vehicle operation shift	Timestamp
9 * 5	Up	1	A	$D0 * 1 *$	118.056	24.6186	Aug. 8, 2018	12	8:00:15
9 * 5	Up	2	B	$D0 * 1 *$	118.058	24.6145	Aug. 8, 2018	12	8:06:45
9 * 5	Up	3	C	$D0 * 1 *$	118.060	24.6129	Aug. 8, 2018	12	8:08:55

the radius of $\text{Dis}^{\text{land-use}}$ meters is taken as the research scope in this paper to get the land-use type attraction coefficient around the stop. There are eight land-use types around stops within the study radius [22]: Residential; Commercial and Business Facilities; Administration and Public Services; Industrial, Manufacturing; Logistics and Warehouse; Municipal Utilities; Green Space and Square; Road, Street, and Transportation.

As there are not only urban construction land but also unused land and other types of land in the study area around bus stops, the total area of urban construction land around each station is not necessarily equal. However, the nonurban construction land has less attraction to the daily life of urban residents, so they are not considered here. What is more, the attraction coefficient is determined according to the scale of the city, and there is no significant difference in its values of cities with similar scales.

2.2.4. Smart Card Data. Raw smart card data in most cities in China are illustrated in Table 3, which contains ID, card type, boarding date, transaction time, line number, direction, and bus license plate number. Users tap smart cards only when boarding, but the boarding location and alighting information are not recorded for efficiency [8].

To get the boarding location, bus GPS data help to detect the minimal time interval of the transaction time and the GPS timestamp, and raw smart card data can be identified with the corresponding stop index, name, longitude, and latitude, as shown in Table 4.

For the identification of records' alighting information, it is necessary to identify destinations by the trip chain method firstly, and the rest of the data of unlinked trips are the focus of this paper. In the case of He and Trépanier, for the last trip of the day, the identification rate when destinations are found in the first trip of the next day is 4.17%, which is lower than that when alighting points are found in the first boarding station of the current day [23]. Besides, Cui's research showed that using multisource IC card data can estimate more alighting stops with higher accuracy than single-source IC card data [12]. Thus, the trip train method used in this paper is based on multisource data and two hypotheses mentioned above without destination found in the first trip of the next day. Records without alighting stop determination by the trip chain method are the study objects in this paper.

2.3. Related Works. Methods used to determine alighting stops of unlinked trips can be divided into aggregate and disaggregate models [24]. The aggregate model's research object is the group, and models can determine the alighting passenger flow at each stop but cannot infer the destination of every smart card record accurately. The method based on the allure of bus stop is popular from aggregate models, and every potential alighting stop attraction can be obtained with factors such as trip distance, nature of land use around the stop, and the transferability of each bus site [25, 26]. In contrast, disaggregate models can identify every record's alighting stop, and they are widely used in more researches, such as the estimation of vehicle load profile and other network performance measures [6, 7]. Therefore, in this work, the framework-based method proposed is disaggregate. More methods of disaggregate models are introduced as below.

2.3.1. Probability Methods. Many scholars calculated the alighting probability at each potential stop in many ways, and the one with the highest probability was the estimated destination. Most researchers used records identified with destinations by the trip chain method as historical datasets to get the alighting probability. He and Trépanier constructed spatial probability and temporal probability in a kernel estimation via the probability of discrete variables with continuous variables, which were multiplied to predict the alighting probability at each potential stop, and the one with the highest probability was the estimated destination [23]. Also, they found a suitable threshold for the distance between estimated and observed alighting stops with the changes in accuracy [27].

However, these approaches failed to cover all destinations of records in unlinked trips with a low identification rate [23, 27].

2.3.2. Step-by-Step Method. Some scholars used several fixed methods to infer alighting stops of unlinked trips, and most of these methods were selected from probability methods.

Table 3: Raw smart card records.

ID	Card type	Boarding date	Transaction time	Line no.	Direction	Bus license plate no.
8 * *350	Common	Aug. 8, 2018	8:06:48	9 * 5	Up	*D*0 * 1 *
8 * *170	Common	Aug. 8, 2018	8:06:51	9 * 5	Up	*D*0 * 1 *
8 * *630	Common	Aug. 8, 2018	8:06:54	9 * 5	Up	*D*0 * 1 *

Table 4: Smart card records with boarding stops determination.

ID	Card type	Boarding date	Transaction time	Line no.	Direction	Bus license plate no.	Boarding stop			
							Index	Name	Longitude	Latitude
8 * *350	Common	Aug. 8, 2018	8:06:48	9 * 5	Up	*D*0 * 1 *	2	B	118.058	24.6145
8 * *170	Common	Aug. 8, 2018	8:06:51	9 * 5	Up	*D*0 * 1 *	2	B	118.058	24.6145
8 * *630	Common	Aug. 8, 2018	8:06:54	9 * 5	Up	*D*0 * 1 *	2	B	118.058	24.6145

Hu et al. used individual characteristics for alighting attraction weighting and identified all records' destinations in their case. Firstly, they used the high-frequency stop method, which picked the potential alighting stop with the highest individual boarding frequency as a destination, and the rest of the data were turned to the next step. Secondly, the ratio of boarding passenger flow in each potential stop to all potential stops was calculated, which helped allocate potential destinations to smart card records randomly. However, the results only showed high reliability in a cluster analysis, and they didn't validate the accuracy of alighting stop determination [24]. Li et al. did a similar study, but at the second step, they chose the stop attraction method, which used the stop with the largest boarding passenger flow in the same shift as the destination. But if no one boarded the bus at the last few stops, the attraction coefficients of these stops were all zero, and selecting the stop with the largest probability was impossible. Therefore, few records may not get alighting stops and the recognition rate was difficult to reach 100%. It was worth mentioning that Li's research got the identification accuracy based on the distance and weight between the true and the estimated destinations [28].

Step-by-step method always used several fixed methods for different datasets, which led to poor generalization ability and low accuracy, and some methods were not enough to identify all alighting stops [24, 28].

2.3.3. Machine Learning Algorithms. Machine learning models could find hidden insights and produce reliable decisions based on the learning from historical data, which were applied on bus alighting stop determination in recent research. Yan et al. developed two-step algorithms with KNN, Decision Tree, Random Forest, or other machine learning algorithms to cover all records of unlinked trips [8]. Destinations were detected by each machine learning algorithm with several features such as origin location, boarding time, bus line number, and two features (number of POIs, distribution of POI points) from POI data instead of land-use information. Besides, this research divided passengers into different groups by using K-means clustering to accurately estimate the alighting stop.

However, machine learning algorithms were more complex and difficult to explain, and parameters needed to be determined in advance and their values had a great influence on the results [8]. In the two-step algorithm with KNN, the value of the nearest neighbor sample was usually subjective and lacked an objective basis. In the two-step algorithm with Decision Tree, if there was no reasonable restriction and pruning for the tree's growth, the small probability events in the training dataset would be completely included, which was prone to overfitting, leading to low prediction accuracy. The two-step algorithm with Random Forest contained multiple decision trees and could reduce overfitting to a great extent. The alighting stop was determined by the output results from multiple decision trees, and the number of trees established in this method was needed to be set manually. In large datasets, too many trees would cost too much time and space, while too few trees reduced the accuracy, so it was difficult to get a suitable value for the number of trees in practical application.

From the literature survey, it can be observed that most of the researchers use only one method (or machine learning algorithm) or several fixed methods in sequence to identify the alighting stops of unlinked trips, and some of them cannot determine all destinations. However, which method or combination is the best for a different dataset is hard to determine, and all alighting stops are expected to be identified as possible. In this work, a two-layer stacking framework is proposed to get the appropriate contributions of several methods in a different dataset simultaneously, which has strong generalization ability and can detect all destinations with high accuracy. Moreover, the step-by-step method of Li [28] and the two-step algorithms with KNN, Decision Tree, or Random Forest of Yan's [8] are chosen as comparison methods.

3. Method

A two-layer stacking framework based on alighting stop determination of unlinked trips can be seen in Figure 3. First, five methods in the first layer are described, and they are high-frequency stop method [28], stop attraction method [28], transfer convenience method [25], land-use type attraction method [26], and I-GHSM presented here. Then, the logistic regression model in the second layer is presented. All results from five methods in the first layer and

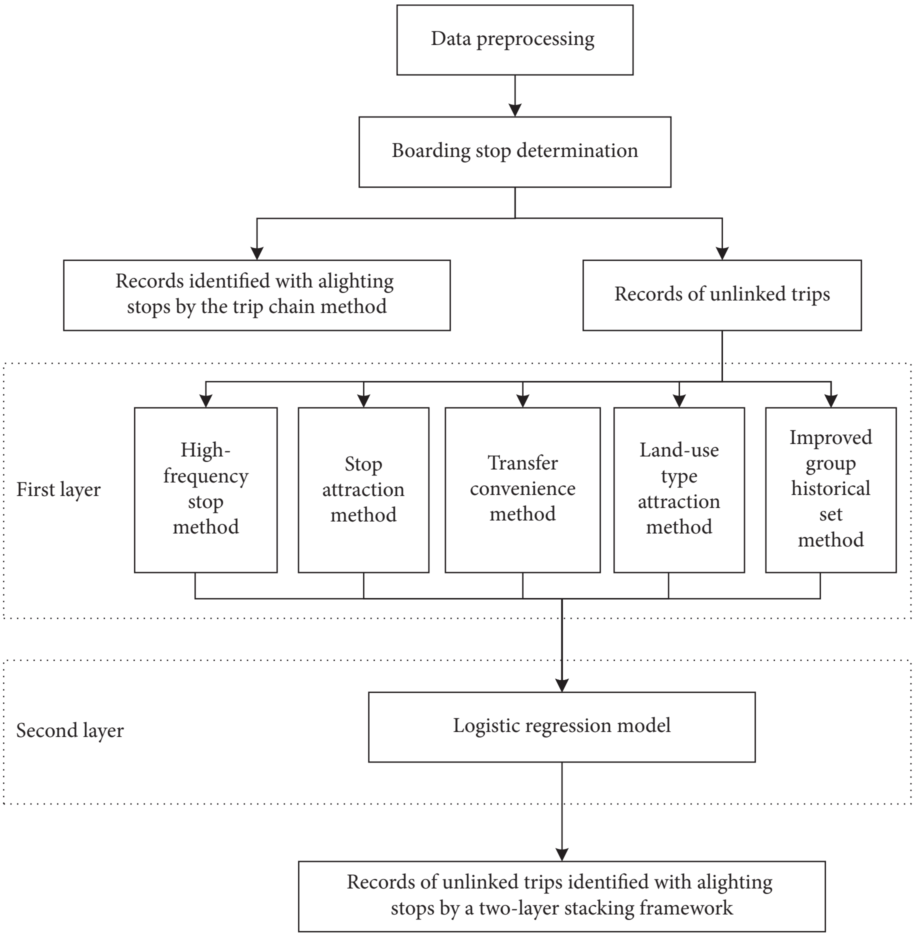
Data preprocessing
Boarding stop determination
Records identified with alighting stops by the trip chain method
Records of unlinked trips
First layer
High-frequency stop method
Stop attraction method
Transfer convenience method
Land-use type attraction method
Improved group historical set method
Second layer
Logistic regression model
Records of unlinked trips identified with alighting stops by a two-layer stacking framework

Figure 3: Alighting stop determination of unlinked trips based on a two-layer stacking framework.

the contribution weight of each method can be obtained via the regression model in the second layer [29]. The two-layer stacking framework uses outputs from the first layer as learning inputs of the second layer, which can correct the systematic deviation in the learning algorithm to improve the accuracy. At last, the destination of an unlinked trip is detected when the probability is maximal.

3.1. Methods in the First Layer. During the research period D, when the passenger whose ID is m boards at the stop $s^{\text{in}}_{m,d,b}$ of trip b on day $d\,(\in D)$, the stop $s^{\text{in}}_{m,d,b}$ is the k_1th stop on direction f of line l, and there are total $K_{l,f}$ stops on the same direction of line l. For the potential alighting stops numbered $k_2\,(k_1 < k_2 \leq K_{l,f})$, alighting probabilities can be obtained by high-frequency stop method [28], stop attraction method [28], transfer convenience method [25], land-use type attraction method [26], and improved group historical set method (I-GHSM), respectively, as shown in Table 5.

3.1.1. High-Frequency Stop Method. There are round-trip characteristics in residents' public transportation [9], which leads to bus stops with higher boarding frequency that have more attraction to alight. For the passenger with ID m, the boarding frequency at every potential stop numbered k_2 is Nup_{k_2} during D days, and the alighting probability of the k_2th stop of trip b can be seen in [28]

$$P^{1}_{m,d,b}(k_1,k_2)=\begin{cases}\dfrac{\mathrm{Nup}_{k_2}}{\sum_{k_2=k_1+1}^{K_{l,f}}\mathrm{Nup}_{k_2}}, & \exists\mathrm{Nup}_{k_2}>0,\\[2ex] 0, & \forall\mathrm{Nup}_{k_2}=0.\end{cases}\tag{1}$$

3.1.2. Stop Attraction Method. Stop attraction means passengers prefer to alight at popular stops with more people getting on the bus. For the trip b of the passenger with ID m, passengers in the same bus shift can be found, and their boarding frequency at stop numbered k_2 is Num_{k_2}. Therefore, the alighting probability of the k_2th stop by the stop attraction method is given as [28]

$$P^{2}_{m,d,b}(k_1,k_2)=\begin{cases}\dfrac{\mathrm{Num}_{k_2}}{\sum_{k_2=k_1+1}^{K_{l,f}}\mathrm{Num}_{k_2}}, & \exists\mathrm{Num}_{k_2}>0,\\[2ex] 0, & \forall\mathrm{Num}_{k_2}=0.\end{cases}\tag{2}$$

Table 5: Five methods in the first layer.

Method	Data source	Alighting probability of the k_2th stop
High-frequency stop method [28]	Smart card data with boarding stops	$P^1_{m,d,b}(k_1, k_2)$
Stop attraction method [28]	Smart card data with boarding stops	$P^2_{m,d,b}(k_1, k_2)$
Transfer convenience method [25]	Static bus stop information	$P^3_{m,d,b}(k_1, k_2)$
Land-use type attraction method [26]	Land-use type attraction coefficient	$P^4_{m,d,b}(k_1, k_2)$
I-GHSM (improved group historical set method)	Smart card data with boarding stops, smart card data determined with destinations by the trip chain method	$P^5_{m,d,b}(k_1, k_2)$

3.1.3. Transfer Convenience Method. If more bus lines are passing through a stop, it is convenient to transfer at this stop, and its attraction is greater than others. There are L_{k_2} bus lines passing through the k_2th stop based on static bus stop information, so the alighting probability of each potential stop by the transfer convenience method can be obtained via equation (3) [25].

$$P^3_{m,d,b}(k_1, k_2) = \frac{L_{k_2}}{\sum_{k_2=k_1+1}^{K_{l,f}} L_{k_2}}. \tag{3}$$

3.1.4. Land-Use Type Attraction Method. Land-use type is one of the determinant factors for bus passengers' destinations. If there are shopping malls near the bus stop, the attraction is greater for people to get off the bus. There are eight types of urban construction land use around each stop, and C_h is the attractive coefficient of the $h \in \{1, 2, \ldots, 8\}$th land-use type. Within the radius of $\text{Dis}^{\text{land-use}}$ meters around the k_2th stop, the land occupation ratio of the hth land-use type is $C_{k_2,h}$. The alighting probability of k_2th stop can be illustrated in equation (4) with the land-use type attraction method [26].

$$P^4_{m,d,b}(k_1, k_2) = \frac{\sum_{h=1}^{H}\left(C_{k_2,h} \cdot C_h\right)}{\sum_{k_2=k_1+1}^{K_{l,f}} \sum_{h=1}^{H}\left(C_{k_2,h} \cdot C_h\right)}. \tag{4}$$

3.1.5. Improved Group Historical Set Method. Bus stops have similar attractions for records with similar behavior patterns. To cluster these records into a group more directly and accurately, based on the GHSM in our previous study, the I-GHSM is proposed here with more indicators considered and using each datum as the smallest research unit. Specific steps of the I-GHSM are described as follows:

(i) Construction of clustering indicators. Several indicators are used to classify records, and there are two types of indicators in this paper [17], as shown in Table 6. In the first type, some fields are picked as base indicators to ensure that the record is the smallest unit. In the second type, we set up some indicators to mine more information. For example, the type of boarding stop is the grade of the passenger flow center to which the station belongs. Records with the same type of boarding stop have similar travel demands. There are three grades in this paper, and they are clustered by the k-means algorithm based on the boarding stop index and its passenger flow. The loyalty of passengers makes records with similar behavior rules closer via clustering the boarding times and the amounts of smart cards into three categories by the *k*-means algorithm. Besides, the *k*-means algorithm mentioned above is described specifically in [30] and [31].

(ii) Normalization of indicators. There are different ranges among eleven indicators in Table 6, so min-max standardization is used to normalize indicators, and the value of each indicator is made to be scaled to unit size.

(iii) Clustering is based on the *K*-means algorithm. The most widely used *K*-means clustering is picked to cluster records, and we can get the optimal number of clustering categories C with groups $\{R_1, \ldots, R_c, \ldots, R_C\}$ via the elbow rule. Data in each group have a similar behavior pattern based on the indicators mentioned above.

(iv) Determining the alighting probability at each possible stop is the last step. All records in the same group R_c with destinations by the trip chain method are used as the group historical dataset. The alighting frequency at every potential stop numbered k_2 is $M^{\text{down}}_{k_2}$ according to the historical dataset, and the alighting probability of the k_2th stop is given as

$$P^5_{m,d,b}(k_1, k_2) = \begin{cases} \dfrac{M^{\text{down}}_{k_2}}{\sum_{k_2=k_1+1}^{K_{l,f}} M^{\text{down}}_{k_2}}, & \exists M^{\text{down}}_{k_2} > 0, \\ \\ 0, & \forall M^{\text{down}}_{k_2} = 0. \end{cases} \tag{5}$$

3.2. Logistic Regression Model in the Second Layer. The logistic regression model with strong interpretability is used in the second layer, which can get each method's contribution weight directly.

3.2.1. Model Constructing. The alighting probabilities at every potential destination are obtained by each method in the first layer, and they are inputs of the logistic regression model. The range of values in the input is [0, 1], so no

Table 6: Indicators of smart card data for clustering.

Type	Index	Indicators	Quantification
Base indicators	1	ID	——
	2	Type	——
	3	Date	Making data into numbers, such as using 20181101 rather than November 1, 2018
	4	Transaction time	Changing transaction time to numbers in seconds with 00 : 00 : 00 a day as a reference, such as using 28800 rather than 8 : 00 : 00
	5	Boarding stop index	——
Constructed indicators	6	Weekday	=0 : The boarding day is not a weekday =1 : The boarding day is a weekday
	7	Weather	=0 : It rains on the boarding day =1 : It doesn't rain on the boarding day
	8	Peak hours	=0 : Transaction time is not in peak hours =1 : Transaction time is in peak hours
	9	Type of boarding stop	=1 : The boarding stop is the first-class passenger flow center =2 : The boarding stop is the second-class passenger flow center =3 : The boarding stop is the third-class passenger flow center
	10	Administrative region of boarding stop	=1 : Boarding stop is in administrative region 1 =2 : Boarding stop is in administrative region 2 The rest can be done in the same manner
	11	Loyalty of passenger	=1 : Bus is an occasional selection =2 : Bus is an alternative selection =3 : Bus is a loyal selection

normalization operation is needed. Each record with a boarding stop is combined with a potential alighting stop to be a pair, and every pair has a label. The label is 1 when the potential destination is the true destination, and 0 stands for other situations.

When the passenger whose ID is m boards at the stop $s^{\text{in}}_{m,d,b}$ on day d of unlinked trip b, the second layer's output is the probability when the label is 1 at the potential k_2th stop in equation (6). Similarly, the probability when the label is 0 can be calculated.

$$P\left(Y_{k_2} = 1 | \vec{x}_{k_2}\right) = \frac{\exp\left(\vec{w}^T \cdot \vec{x}_{k_2}\right)}{1 + \exp\left(\vec{w}^T \cdot \vec{x}_{k_2}\right)}, \tag{6}$$

where $\vec{x}_{k_2} = [P^1_{m,d,b}(k_1, k_2), P^2_{m,d,b}(k_1, k_2), P^3_{m,d,b}(k_1, k_2), P^4_{m,d,b}(k_1, k_2), P^5_{m,d,b}(k_1, k_2), 1]^T$ is the input vector; $\vec{w} = [w_1, w_2, w_3, w_4, w_5, b]^T$ is the weight vector to be learned, where b is bias.

3.2.2. Model Learning. When learning the model, records identified with alighting stops by the trip chain method based on multisource data are used as the training and testing datasets. Records of unlinked trips can be determined with destinations via the learned model.

Every boarding record only has one alighting stop, and records with the label 0 are more than that with label 1. To solve the problem that the smart card data is unbalanced, the large sample data labeled 0 is sampled randomly, and the small sample labeled 1 is sampled based on the SMOTE algorithm. Then the two parts of data merge as a new dataset, and 90% of the set is selected randomly as the training set, while the remaining 10% is the testing set to verify the validity of the model with the evaluation index of F1-Score.

The maximum likelihood estimation method is used to estimate the parameters, and the Limited-memory BFGS (L-BFGS) algorithm for large-scale data computing is chosen to solve the optimization problem aiming to maximize the likelihood function. When the maximum likelihood estimation of $\vec{w}$ is $\hat{\vec{w}}$, we can get the probability of k_2th stop when the label is 1 as in equation (7), and the probability when the label is 0 can be calculated similarly.

$$P\left(Y_{k_2} = 1 | \vec{x}_{k_2}\right) = \frac{\exp\left(\hat{\vec{w}}^T \cdot \vec{x}_{k_2}\right)}{1 + \exp\left(\hat{\vec{w}}^T \cdot \vec{x}_{k_2}\right)}. \tag{7}$$

3.3. Alighting Stop Determination of Unlinked Trips. To ensure that every boarding record has one alighting stop, we choose the potential alighting stop with the maximum probability when the label is 1 as the destination.

When the passenger whose ID is m boards at the stop $s^{\text{in}}_{m,d,b}$ on day d of unlinked trip b, the alighting stop is the k_2th stop as shown in equation (8). Furthermore, alighting stop name, longitude, and latitude can be obtained by combining the stop index with static bus stop information.

$$k_2 = \arg\max_{k_2}\left\{P\left(Y_{k_1+1} = 1|\overrightarrow{x}_{k_2}\right), \ldots, P\left(Y_{k_2} = 1|\overrightarrow{x}_{k_2}\right), \ldots, P\left(Y_{K_{l,f}} = 1|\overrightarrow{x}_{k_2}\right)\right\}. \tag{8}$$

4. Case Analysis

4.1. Dataset. Xiamen BRT (Bus Rapid Transit System) is the first BRT system adopting the viaduct mode in China. The smart card data of BRT contains complete information of boarding and alighting stops, and the physical isolation of lanes makes several BRT lines form a small bus network. Smart card records with boarding and alighting stops along BRT Line Kuai 1 in Xiamen are research subjects in this paper. Besides, data of other BRT lines in Xiamen are used to complete passengers' trip trains (see Figure 4).

Selecting data of IC card owners who are on BRT Line Kuai 1 in November 2018, 3673184 records with boarding stops are obtained. Among them, 2425101 records can be identified with alighting stops by the trip chain method with multisource data, and the accuracy is 80.96%, as shown in Figure 5 [12]. Based on all records with boarding stops, more than 80% of passengers traveled only 8 times or less this month, so the individual historical set (the card history) is small, and the group historical set method proposed is needed. Also, there are four types of IC cards including common cards, student cards, elderly cards, and special cards. The morning peak is 7 : 00 : 00–9 : 00 : 00 and the evening peak is from 17 : 00 : 00–19 : 00 : 00 in November 2018, and the weather in each day can be obtained from the China Meteorological Administration.

BRT Line Kuai 1 has 27 stops, and its up and down directions run through the same stops. These stops are distributed in three administrative areas: Siming District, Huli District, and Jimei District. The land-use type distributions around 27 stops within 800 meters are illustrated in Figure 6. The attraction coefficient varies because of the different land-use nature, but it has good portability in similar cities. This paper sets the attraction coefficient of each land-use type in Table 7, according to experiences and relevant studies [32]. What's more, the number of lines passing through each stop along BRT Line Kuai 1 can be counted with static bus stop information and is shown in Figure 7.

4.2. Evaluation Methods. The identification rate Iden and the accuracy Acc are used as evaluation indexes to measure the method's performance, which is shown in equations (9) and (10), respectively.

$$\text{Iden} = \frac{N_{\text{iden}}^{\text{un}}}{N^{\text{un}}}, \tag{9}$$

$$\text{Acc} = \frac{N_{\text{iden}-r}^{\text{un}}}{N_{\text{iden}}^{\text{un}}}, \tag{10}$$

, where N^{un} is the total number of records that need to be identified for alighting stops in unlinked trips, $N_{\text{iden}}^{\text{un}}$ is the number of records that can be determined with destinations, and $N_{\text{iden}-r}^{\text{un}}$ is the number of records with destinations predicted correctly. The higher the values of Iden and Acc, the higher the prediction accuracy of the model.

4.3. Parameters Setting. Parameters involved in this case are determined as follows:

4.3.1. Distance Threshold Setting Based on the Trip Chain Method. According to the distance between BRT stops in Xiamen, this paper sets the radius of $\text{Dis}^{\text{land- use}} = 800$ meters and the distance threshold in the trip chain method as 2000 meters.

4.3.2. Determination of Penalty Coefficient in the Logistic Regression Model. In the second layer of our method, after many experiments, the best value of penalty coefficient in the logistic regression model is 100.

4.3.3. Parameters Setting of Comparison Methods. Two-step algorithms with KNN, Decision Tree, Random Forest [8], and the step-by-step method of Li [28] are selected for comparison. The theoretical analysis and comparison between these methods and the method proposed have been introduced in Section 3.1.3.

In two-step algorithms with KNN (K-Nearest Neighbor), Decision Tree, and Random Forest, existing studies use POI data to replace the land-use type around stops. However, the land-use type is known in this case, so there is no need to use POI data instead. After many experiments, it is determined that the nearest neighbor sample value is 1000 in the KNN, the number of trees established is 2000 in the two-step algorithms with Random Forest, and the Gini coefficient is selected as the standard in the two-step algorithm with Decision Tree.

The step-by-step method of Li [28] and our approach all need to use the passenger flow at each stop of the same bus shift. However, BRT's smart card records only have boarding and alighting stops, and we cannot sort out the shift passengers took. Therefore, each station's passenger flow within the hour of the passenger's timestamp is taken as the passenger flow needed in this case.

4.4. Results

4.4.1. Methods in the First Layer. In the first layer, I-GHSM is proposed, whose results of clusters are introduced as given below. Besides, the comparisons of I-GHSM and GHSM in the method based on expanded history trip records are shown. At last, results of unlinked trips' destinations determined by each method in the first layer are calculated.

In I-GHSM, when constructing clustering indicators, the passenger loyalty index should be calculated firstly, which is

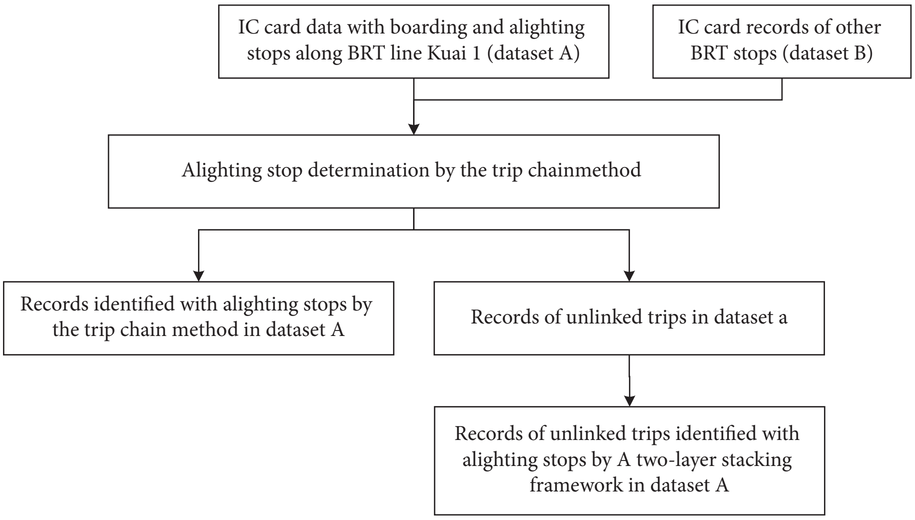
IC card data with boarding and alighting stops along BRT line Kuai 1 (dataset A)
IC card records of other BRT stops (dataset B)
Alighting stop determination by the trip chainmethod
Records identified with alighting stops by the trip chain method in dataset A
Records of unlinked trips in dataset a
Records of unlinked trips identified with alighting stops by A two-layer stacking framework in dataset A

Figure 4: Schematic diagram of data relations used in this case.

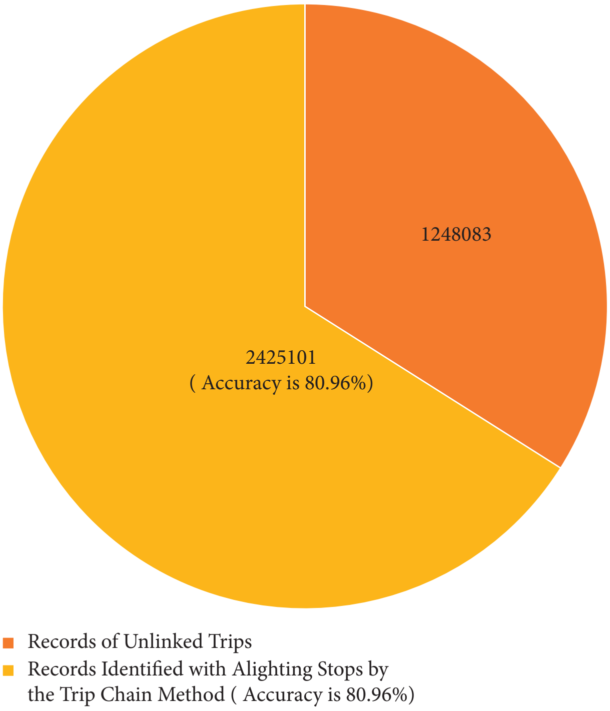

Figure 5: The relationship of smart card data in this case.

based on clustering the boarding times and the number of smart cards into three categories with the k-means algorithm. In this case, when the boarding frequency is 1 or 2 times per month, the passenger occasionally takes BRT with the value of passenger loyalty being 1; when the boarding frequency is 3 to 8 times per month, the passenger occasionally takes BRT as an alternative trip mode with the value being 2; when the boarding frequency is more than 8 times per month, the passenger is a loyal user of BRT and the value is 3. All eleven indicators in the I-GHSM can be determined and then processed by min-max standardization. After that, K-means clustering is done to cluster records, and the variation of average distortion degree with the number of clusters is obtained, as shown in in Figure 8. As a result, the best number of clusters is 2 due to the elbow rule, and 3673184 records are divided into clusters R_1 and R_2, as shown in Table 8.

In the method based on expanded history trip records, the trip chain method with multisource data is used to estimate alighting stops firstly. For records in the unlinked trips, the individual historical set method and the GHSM are used in sequence for estimating alighting stops. However, using the I-GHSM as the third method can get better results, as shown in Table 9.

There are five methods in the first layer: the high-frequency stop method [28], the stop attraction method [28], the transfer convenience method [25], the land-use type attraction method [26], and the I-GHSM. If the potential alighting stop with the maximum probability is the destination, each method's results of alighting stop determination of unlinked trips are shown in Table 10.

4.4.2. Logistic Regression Model in the Second Layer. In this part, 2425101 records detected with alighting stops by the trip train method and multisource data help to learn and test the logistic regression. In the model training, each datum's boarding stop should be combined with each potential alighting stop to be a new pair, and there are a total of 40113752 new pairs in this case. Among them, 2425101 pairs' labels are 1 with true destinations, and the rest 37688651 pairs' labels are 0. Each pair can be considered as a new complete record. Considering the imbalance of data with different labels, records are resampled, as shown in Table 11.

To prevent the model from underfitting caused by the training dataset being too small, the data with label 0 after sampling is still more than expected, as seen in Table 10. However, the data ratio (Label is 0: Label is 1) has been reduced from 15.54 : 1 to 1.82 : 1 after sampling, which is a good improvement.

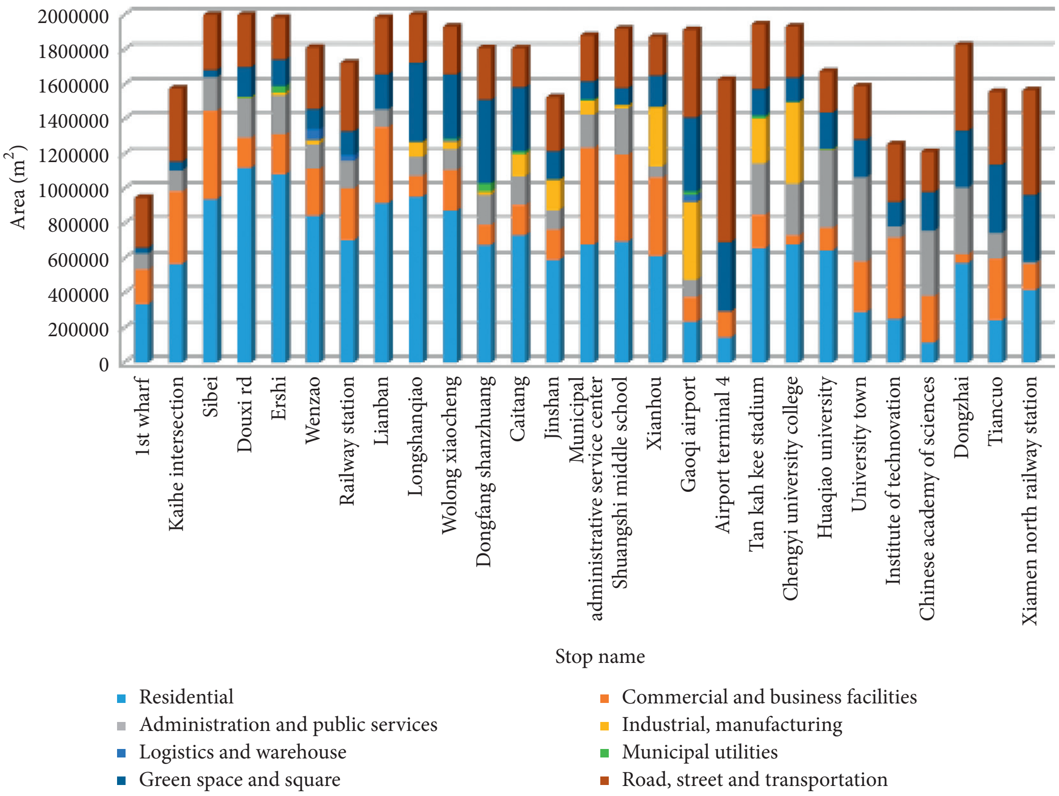

Figure 6: Land-use types and area coverage within 800 meters around stops along Xiamen BRT Line Kuai 1.

Table 7: Attractive coefficient of various land-use types.

Land-use type	Attractive coefficient
Residential	1
Commercial and business facilities	1.2
Administration and public services	1.1
Industrial, manufacturing	1
Logistics and warehouse	0.6
Municipal utilities	0.8
Green space and square	0.7
Road, street, and transportation	1.3

For 75178131 records after sampling, 90% of them are randomly selected as the training set of the logistic regression model and the remaining 10% of them as the test set. When the penalty coefficient is 100, the F_1-Score is 0.67, and the parameters of the trained logistic regression model are shown as equation (11).

$$\vec{w} = [5.7141, 0.0769, 0.1527, -0.1713, 0.4785, 2.4893]^T. \quad (11)$$

4.4.3. Alighting Stop Determination of Unlinked Trips. For 1248083 records of unlinked trips, identification rate and accuracy of existing methods [8, 28] and the two-layer stacking framework method proposed are in Table 12.

The accuracy of destination recognition based on a two-layer stacking framework changes in different conditions, and it can be calculated as follows:

(1) According to different days of one week in our study period, passenger transaction time can be divided into hours, and the accuracy of each period is shown in Figure 9

(2) According to different card types, passenger transaction time can be divided into hours, and the accuracy of each period is shown in Figure 10

(3) According to different card types and loyalty, each division's accuracy rate is counted and shown in Figure 11

4.5. Results Analysis.

(1) According to Table 9, the accuracy of unlinked trips' alighting stop determination with the individual historical set method and I-GHSM is 33.20%, which is 7.83% higher than that of the previous method. Therefore, I-GHSM achieves the goal that records with similar behavior patterns are clustered into a group more accurately than GHSM.

(2) As can be seen from Tables 10 and 12, the identification rate or the accuracy of each method in the first layer is lower than that of the two-layer stacking framework method. These verify that the contribution weight of each method obtained by the logistic regression model in the second layer is suitable enough to get better results, which brings the generalization ability for this dataset.

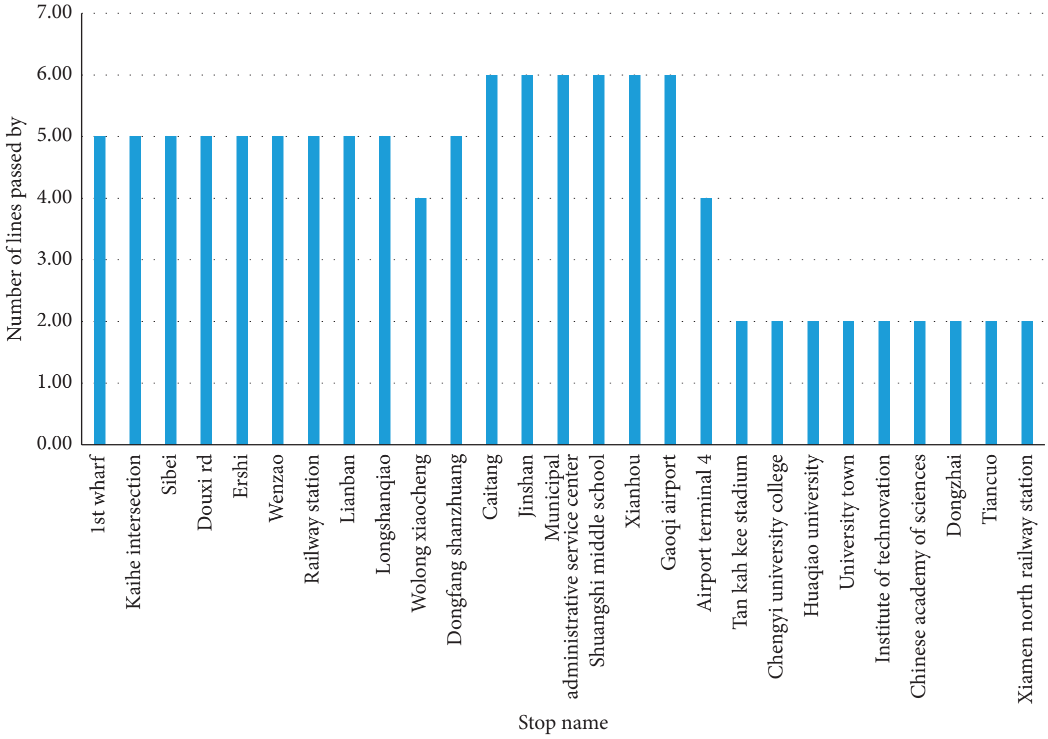

Figure 7: The stop name and the number of lines passed by.

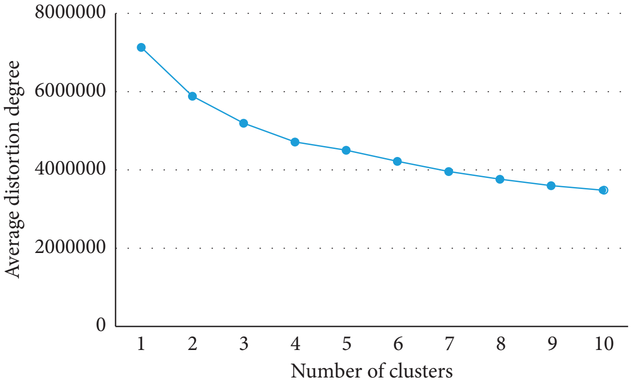

Figure 8: The average distortion degree varies with the number of clusters.

Table 8: Number of records in different clusters.

Cluster	Records with boarding stops	Records determined with alighting stops by the trip chain method	Records of unlinked trips
R_1	2129790	1351757	778034
R_2	1543394	1073344	470049
Total	3673184	2425101	1248083

(3) In Table 12, the accuracy of the two-layer stacking framework method is the highest than comparison methods, which indicates that using several models simultaneously is more effective than choosing only one or several fixed methods in sequence. Besides, the two-layer stacking framework-based method detects all alighting points of unlinked trips, which is better than the step-by-step method of Li [28].

(4) As shown in Figure 9, the accuracy obtained during the morning peak (7 : 00 : 00–9 : 00 : 00) on weekdays was higher than that on weekends. In Figure 10, records from common and student cards account for

Table 9: Results of unlinked trips' alighting stop determination with the GHSM or the I-GHSM.

Method	Records identified with alighting stops	Identification rate	Records identified with alighting stops correctly	Accuracy (%)
Individual historical set method	446915	35.81	273628	61.23
GHSM/**I-GHSM**	801168	64.19	43042/**140721**	5.37/**17.56**
Total	1248083	100.00	316670/**414349**	25.37/**33.20**

Table 10: Results of unlinked trips' alighting stop determination by each method in the first layer.

Method	Records identified with alighting stops	Identification rate	Records identified with alighting stops correctly	Accuracy
High-frequency stop method [28]	926202	74.21	582581	62.90
Stop attraction method [28]	1248083	100.00	146400	11.73
Transfer convenience method [25]	1248083	100.00	118818	9.52
Land-use type attraction method [26]	1248083	100.00	141907	11.37
I-GHSM	1248083	100.00	225653	18.08

Table 11: The number of records before and after data sampling.

Type		Label is 0	Label is 1	Total
Before sampling		37688651	2425101	40113752
After sampling	Random under-sampling	24251010	2425101	75178131
	Oversampling based on SMOTE algorithm	24251010	24251010	

Table 12: Results of alighting stop determination of unlinked trips in different methods.

Method	Records identified with alighting stops	Identification rate (%)	Records identified with alighting stops correctly	Accuracy (%)
Two-step algorithms with KNN [8]	1248083	100.00	124808	10.00
Two-step algorithms with decision tree [8]	1248083	100.00	160379	12.85
Two-step algorithms with random forest [8]	1248083	100.00	202689	16.24
Step-by-step method of Li [28]	1247536	99.96	623876	50.01
Two-layer stacking framework method	**1248083**	**100.00**	**647538**	**51.88**

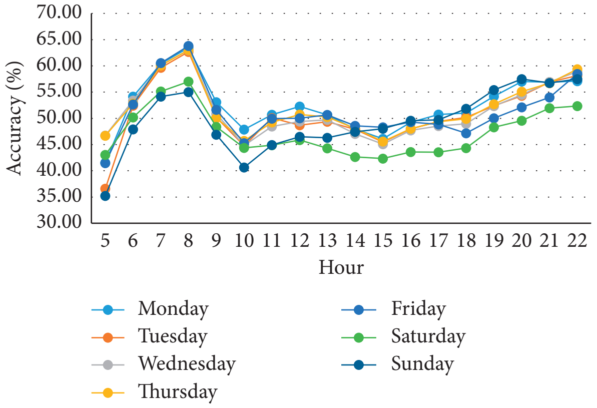

Figure 9: The accuracy based on the method proposed in different hours and days of the week.

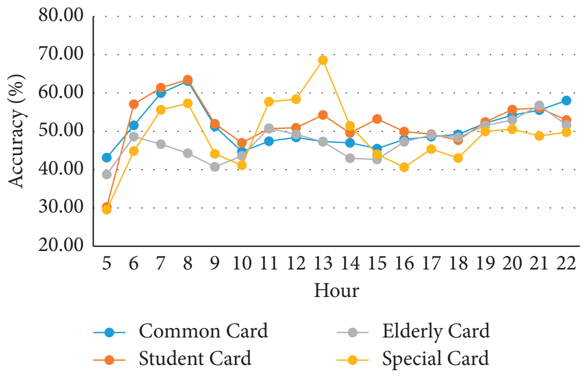

FIGURE 10: The accuracy based on the method proposed in different hours and card types.

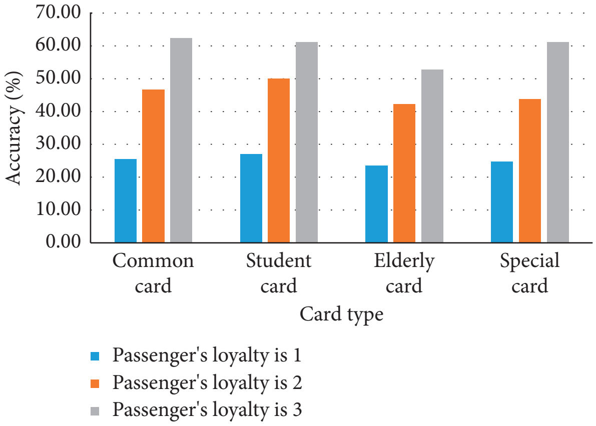

FIGURE 11: The accuracy based on the method proposed in different card types and loyalty.

more than 93.5% of the total smart card data, and their accuracy in the morning peak is higher than that in other periods. Therefore, the morning peak on weekdays is the most regular time for passengers to travel by public transport, which is consistent with the fact that there are many commuters (/students) from their fixed residence to their fixed workplace (/school) during this period. From the analysis of Figure 11, all card types meet the condition that higher boarding frequency leads to higher accuracy of alighting stops determination, which shows that high loyalty passengers' travel behavior is more regular than that of other passengers with lower loyalty.

5. Conclusions

In this paper, a method based on a two-layer stacking framework is proposed to get better accuracy in unlinked trips' alighting stop determination, and the improved group historical set method is presented in the first layer.

Xiamen's case shows that the I-GHSM can cluster records with similar behavior patterns into a group more accurately than the GHSM. Together with the individual historical set method, the I-GHSM improves accuracy by 7.83% in unlinked trips' destination determination than our previous study. Besides, the method based on the two-layer stacking framework can detect all alighting points with a higher accuracy of 51.88%, which is better than the step-by-step method of Li [28], and two-step algorithms with KNN, Decision Tree, or Random Forest of Yan [8] in this case.

When comparing the identification results of each method in the first layer and the two-layer stacking framework method, the logistic regression model in the second layer is verified to bring the appropriate contribution weight of each method and the generalization ability. And, because the accuracy of the two-layer stacking framework method is the highest than comparison methods, using several models simultaneously is more effective than choosing only one or several fixed methods in sequence. After analyzing changes in identification accuracy based on the framework-based method in different conditions, the morning peak on weekdays is the most regular time for passengers to travel. And, the higher the boarding frequency is, the more regular their travel behavior will be.

However, in the second layer of the two-layer stacking-framework-based method, smart card data are still unbalanced after sampling when determining alighting stops of unlinked trips. Therefore, the sampling method needs to be further studied. What's more, our method can be applied to different datasets of other cities in the future to further verify its effectiveness.

Data Availability

The data used to support the findings of this study are available in Cui, Z., Wang, C., Chen, D., Lei, L. "Alighting stop determination of transit passengers based on expanded history trip records". Journal of Nanjing University (Natural Science), 2020, 56(2), 227–235. https://doi.org/10.13232/j.cnki.jnju.2020.02.008.

Conflicts of Interest

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

Acknowledgments

Ziwei Cui, Cheng Wang, and Yueer Gao contributed equally to this work. This research was supported by the National Natural Science Foundation of China (Grant Nos. 51608209 and 71571056), the National Key Technology R&D Program of China (Grant No. 2018YFB1402500), the Natural Science Foundation of Fujian Province (Grant No. 2017J01090), the Guiding Project of Fujian Science and Technology Plan (Grant No. 2019H0017), the Social Science Planning Foundation of Fujian Province (Grant No. FJ2020B0033), the Project of science and technology plan of Fujian Province of China (Grant No. 2020H0016), the Project of Quanzhou Science and Technology Plan (Grant No. 2018Z008), the Scientific Research Funds of Huaqiao University (Grant No. 16BS304), the Project of Postgraduate Research Innovation Ability Cultivation of Huaqiao

University (Grant No. 17013083017), and the Project of Postgraduate Research Innovation Ability Cultivation of Huaqiao University (Grant No. 17013083024).

References

[1] M. Bagchi and P. R. White, "The potential of public transport smart card data," *Transport Policy*, vol. 12, no. 5, pp. 464–474, 2005.

[2] 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.

[3] C. Morency, M. Trépanier, and B. Agard, "Measuring transit use variability with smart-card data," *Transport Policy*, vol. 14, no. 3, pp. 193–203, 2007.

[4] F. Devillaine, M. Munizaga, and M. Trépanier, "Detection of activities of public transport users by analyzing smart card data," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2276, no. 1, pp. 48–55, 2012.

[5] 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.

[6] M. Trépanier and F. Vassiviere, "Democratized smartcard data for transit operator," in *Proceedings of the 15th World Congress on Intelligent Transport Systems and ITS America's 2008 Annual MeetingITS AmericaERTICOITS JapanTransCore*, New York, NY, USA, 2008.

[7] M. Trépanier, C. Morency, and B. Agard, "Calculation of transit performance measures using smartcard data," *Journal of Public Transportation*, vol. 12, no. 1, pp. 79–96, 2009.

[8] F. Yan, C. Yang, and S. V. Ukkusuri, "Alighting stop determination using two-step algorithms in bus transit systems," *Transportmetrica A: Transport Science*, vol. 15, no. 2, pp. 1522–1542, 2019.

[9] W. Liu, Q. Tan, and L. Liu, "Destination estimation for bus passengers based on data fusion," *Mathematical Problems in Engineering*, vol. 2020, Article ID 8305475, 10 pages, 2020.

[10] M. A. Munizaga and C. Palma, "Estimation of a disaggregate multimodal public transport origin-destination matrix from passive smartcard data from Santiago, Chile," *Transportation Research Part C: Emerging Technologies*, vol. 24, pp. 9–18, 2012.

[11] Q. Zou, X. Yao, P. Zhao, H. Wei, and H. Ren, "Detecting home location and trip purposes for cardholders by mining smart card transaction data in Beijing subway," *Transportation*, vol. 45, no. 3, pp. 919–944, 2018.

[12] Z. Cui, C. Wang, D. Chen, and L. Lei, "Alighting stop determination of transit passengers based on expanded history trip records," *Journal of Nanjing University (Natural Science)*, vol. 56, no. 2, pp. 227–235, 2020.

[13] J. L. Machado, R. De Oña, F. Diez-Mesa, and J. De Oña, "Finding service quality improvement opportunities across different typologies of public transit customers," *Transportmetrica A: Transport Science*, vol. 14, no. 9, pp. 761–783, 2018.

[14] A. Bhaskar and E. Chung, "Passenger segmentation using smart card data," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 3, pp. 1537–1548, 2014.

[15] C. Yang, F. Yan, and S. V. Ukkusuri, "Unraveling traveler mobility patterns and predicting user behavior in the Shenzhen metro system," *Transportmetrica A: Transport Science*, vol. 14, no. 7, pp. 576–597, 2018.

[16] X. Ma, C. Liu, H. Wen, Y. Wang, and Y.-J. Wu, "Understanding commuting patterns using transit smart card data," *Journal of Transport Geography*, vol. 58, pp. 135–145, 2017.

[17] K. Mohamed, E. Côme, L. Oukhellou, and M. Verleysen, "Clustering smart card data for urban mobility analysis," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 3, pp. 712–728, 2016.

[18] P. Kumar, A. Khani, and Q. He, "A robust method for estimating transit passenger trajectories using automated data," *Transportation Research Part C: Emerging Technologies*, vol. 95, pp. 731–747, 2018.

[19] N. Nassir, M. Hickman, and Z.-L. Ma, "Activity detection and transfer identification for public transit fare card data," *Transportation*, vol. 42, no. 4, pp. 683–705, 2015.

[20] A. A. Nunes, T. G. Dias, and J. F. 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, 2015.

[21] G. E. Sánchez-Martínez, "Inference of public transportation trip destinations by using fare transaction and vehicle location data," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2652, no. 1, pp. 1–7, 2017.

[22] Chinese Standard Net, *Code for Classification of Urban Land Use and Planning Standards of Development Land: GB 50137-2011*, Chinese Standard Net, Beijing, China, 2010.

[23] L. He and M. Trépanier, "Estimating the destination of unlinked trips in transit smart card fare data," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2535, no. 1, pp. 97–104, 2015.

[24] J. Hu, J. Deng, and Z. Huang, "Trip-chain based probability model for identifying alighting stations of smart card passengers," *Journal of Transportation Systems Engineering and Information Technology*, vol. 14, no. 2, pp. 62–67, 2014.

[25] M. Zhang, Y. Guo, and Y. Ma, "A probability model of transit OD distribution based on the allure of bus station," *Journal of Transport Information and Safety*, vol. 3, pp. 57–61, 2014.

[26] W. Xu, C. Deng, and B. Liu, "Approach on public traffic passenger flow statistics based on IC data," *China Journal of Highway and Transport*, vol. 26, no. 5, pp. 158–163, 2013.

[27] L. He, N. Nassir, M. Trépanier, and M. Hickman, *Validating and Calibrating a Destination Estimation Algorithm for Public Transpot Smart Card Fare Collection Systems*, CIRRELT), Montreal, Canada, 2015.

[28] J. Li, J. Zhang, J. Zhang, Q. Wang, and D. Peng, "An algorithm to identify passengers' alighting stations and the effectiveness evaluation," *Geomatics and Information Science of Wuhan University*, vol. 43, no. 8, pp. 1172–1177, 2018.

[29] D. H. Wolpert, "Stacked generalization," *Neural Networks*, vol. 5, no. 2, pp. 241–259, 1992.

[30] X. S. Lu, M. Zhou, L. Qi, and H. Liu, "Clustering-algorithm-based rare-event evolution analysis via social media data," *IEEE Transactions on Computational Social Systems*, vol. 6, no. 2, pp. 301–310, 2019.

[31] X. Xu, J. Li, M. Zhou, J. Xu, and J. Cao, "Accelerated two-stage particle swarm optimization for clustering not-well-separated data," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 50, no. 11, pp. 4212–4223, 2020.

[32] Q. Xiao and W. Xu, *Urban Traffic Planning*, pp. 212–282, China Communications Press Co., Ltd, Beijing, China, 1990.