Hindawi
Complexity
Volume 2018, Article ID 3412070, 14 pages
https://doi.org/10.1155/2018/3412070

Research Article

A Trip Purpose-Based Data-Driven Alighting Station Choice Model Using Transit Smart Card Data

Kai Lu,[1] Alireza Khani,[2] and Baoming Han[1]

[1] *School of Traffic and Transportation, Beijing Jiaotong University, Beijing 100044, China*
[2] *Department of Civil, Environmental and Geo-Engineering, University of Minnesota, Minneapolis, MN 55455, USA*

Correspondence should be addressed to Kai Lu; lukai_bjtu@163.com and Baoming Han; bmhan@bjtu.edu.cn

Received 18 December 2017; Revised 2 June 2018; Accepted 15 July 2018; Published 28 August 2018

Academic Editor: Shuliang Wang

Automatic fare collection (AFC) systems have been widely used all around the world which record rich data resources for researchers mining the passenger behavior and operation estimation. However, most transit systems are open systems for which only boarding information is recorded but the alighting information is missing. Because of the lack of trip information, validation of utility functions for passenger choices is difficult. To fill the research gaps, this study uses the AFC data from Beijing metro, which is a closed system and records both boarding information and alighting information. To estimate a more reasonable utility function for choice modeling, the study uses the trip chaining method to infer the actual destination of the trip. Based on the land use and passenger flow pattern, applying k-means clustering method, stations are classified into 7 categories. A trip purpose labelling process was proposed considering the station category, trip time, trip sequence, and alighting station frequency during five weekdays. We apply multinomial logit models as well as mixed logit models with independent and correlated normally distributed random coefficients to infer passengers' preferences for ticket fare, walking time, and in-vehicle time towards their alighting station choice based on different trip purposes. The results find that time is a combined key factor while the ticket price based on distance is not significant. The estimated alighting stations are validated with real choices from a separate sample to illustrate the accuracy of the station choice models.

1. Introduction

In the late 1990s, smartcard payment systems were installed in some big cities, and after more than twenty years of development, more than one hundred cities over five continents have adopted smartcard payment systems [1]. This technology has become pivotal to ticket fare collection for public transit for both bus and metro. Since its inception, the transit smart card system produced a large amount of very detailed data on on-board transactions [2]. The smart data system contains many aspects such as the hardware technology (radiofrequency identification (RFID), electromagnetic shield), system construction, and data storage [3, 4]. Meanwhile, with the rich data source data collected by smart card, a lot of researchers are interested in the applications of those data. Generally, the application can be classified into three levels: strategic level, tactical level, and operational level [5]. For the strategic level, the large amount of data from smart card gives an opportunity for tracking and analyzing long-term individual travel behaviors in both spatial and temporal dimensions. The valuable historical data are fundamental data input for short-term or long-term transit network planning [6]. At the same time, tracking the starting and ending date for each user could obtain the life span of each transit user, which is the supplemental input for network planning [7]. Tactical level is the research related to the strategies that are trying to improve the efficiency, benefits, and energy consumption of the transit system [8]. Operational level is the most popular topic in data application. Generally, there are two branches in this research: passenger behavior analysis and service adjustments. We believe that based on travel information recorded by smart card data, the passenger behavior such as route choice and transfer station choice during their journey in the transit network can be deduced [9–11]. In order to provide better service for the passengers and save their travel time, the timetables are rescheduled based on the variable passenger demand [12]. Meanwhile, operation agencies could estimate and evaluate the transit service performance by operational statistics such as bus run time, vehicle-kilometers, and person-kilometers [13–16].

In addition to the closed travel information loop for each transit user, in some transit systems, passengers are required to tap the card only while they enter the vehicle, which provide only boarding information [17]. In these systems, the one of the boarding or alighting information is missing. Thanks to the automated vehicle location (AVL), automated data collection (ADC), and other support data resource, merging various transit datasets makes it possible to complete the travel route information. For the past 10 years, researchers worked on finding the closed information for each individual trip. Table 1 summarizes the literatures on seeking missing information in open systems including the methodologies, pros, and future research. In Table 1, AFC is short for automatic fare collection. ADCS is short for automated data collection systems, and AVL is short for automatic vehicle location.

Trip chaining methodology is the typical methodology in these research. Here are two basic assumptions: (1) A high percentage of riders return to the destination station of their previous trip to begin their next trip, and (2) a high percentage of riders end their last trip of the day at the station where they began their first trip of the day. In addition to applying the basic assumptions, for each cardholder, there should be more than one trip in the system. Otherwise, it is impossible to infer the alighting station. For some passengers such as commuters, multiday travel information is recorded. The single trip destination could be inferred based on records from other days. If there is only a one-day trip for the cardholder and contains only one trip, the alighting station is invalid.

For passengers, when choosing the alighting station, they consider the in-vehicle time, transfer time, walking time, and ticket fare comprehensively and choose the station which has the highest utility. Sometimes, the alighting station differs based on different trip purposes because the time value could vary for different purposes. To formulate this optimization model, it is necessary to validate the weight and the coefficient for those impact parameters. Because of the missing information and lack of closed trip data, the validation of those models is seldom discussed. The early attempt to validation and sensitivity analysis is based on the on-board survey data to illustrate the feasibility of the method. However, the on-board survey is expensive and data samples are limited.

The Beijing metro system is a closed system, which contains both boarding and alighting information. With walking time, in-vehicle time, and ticket fare for each candidate alighting a station in a buffer walking time for each trip and the real alighting station from AFC data, the coefficient of each utility factor is estimated. Inspired by Tavassoli et al. [28], we relaxed the alighting station information in AFC data from the Beijing metro system to estimate the alighting station for the different trip purposes to see what choice model could illustrate passenger behavior based on different trip purposes. The choice model calibration results for the different trips could be used for passenger behavior analysis, network planning, and policy applications.

This paper is organized as follows. In the following section, it describes the data and data preparation process. In the next section, the method for determining trip purposes, trip origins, and trip destinations is presented. In the methodology section, a multinomial logit model and mixed logit models with independent and correlated normally distributed random coefficients are proposed. We used the AFC data to calibrate the parameters in different models in the first and second parts of the empirical study. In the last part of the empirical study, a separate sample of AFC records is used to illustrate the model's accuracy and validity. Conclusions and directions for future work are presented in the last section.

2. Data: Beijing Metro Transit

2.1. Data Description. The data used in this paper are obtained from a metro transit in Beijing, China, and were excerpted from one week of data, in December 2016. At that time, there were 17 lines serving more than 10 million passengers every day with more than 8000 train services. The majority of line headways ranged from 2 to 5 min, and in the peak hour, the headway could reach 90 s. There are two kinds of payment in Beijing metro, a Yikatong card, which can be charged and used for several times, and one trip pass. The proportion of the Yikatong cardholder among all transit passengers is roughly 80%, and only the Yikatong card data can be recorded in the AFC system. In this research, the AFC data, station geometry data, and timetable data are required, and Table 2 represents the data recorded in the dataset.

The AFC dataset contains the entry and exit information for each passenger. One record represents a trip for a passenger. For example, a passenger started his trip from Xizhimen Station at 8 : 00 AM and alighted at Dongzhimen Station at 8 : 30 AM. Every station has a unique station ID and station location. For a normal station, the route ID saved only one route. For a transfer station, it serves more than one route, so the route ID contains more than one route. For example, Xizhimen Station is a transfer station for route 13, route 2, and route 4. This station only has one unique station ID, station name, and station location in the dataset. The 3 routes are saved in the route ID. The timetable dataset recorded the train arrival and departure time at each stop for each route. The passenger in-vehicle time could be inferred. In Beijing, the ticket price is based on the shortest travel distance and does not take route into consideration. For example, one passenger started his trip from Xizhimen Station to Dongzhimen Station; regardless of whether he takes route 13 or route 2, the ticket price is the same.

In the database discussed above, the AFC data provide the sample for the empirical study. Walking distances were calculated as the Euclidean distance, and the timetable was used to calculate the travel time between stations using the shortest path.

Table 1: Review of studies on estimating alighting stop in a tap-in transit system.

Author	Data	Assumption and constraints	Analysis/use methodology	Application	Pros	Limitations
Barry et al. (2002) [18]	AFC	Two basic assumptions	Trip chaining	New York	Easy to apply	Lack of one trip estimation
Zhao et al. [19] (2007) and Zhao (2004) [20]	ADC	Walking distance threshold	Database management systems	Chicago	(i) Integrating the AFC and AVL (ii) examining the spatial connection	The model was just focused on the bus and rail station
Trépanier et al. (2007) [21]	AFC	Walking tolerance is 2 km.	Transportation object-oriented modeling with vanishing route set	Gatineau	The model is quite suitable for regular transit users	Some passenger information such as single ticket user is missing.
Chu and Chapleau (2008) [22]	AFC	5 min temporal leeway for uncertainty	The linear interpolation and extrapolation to infer the vehicle position	Société de transport de l'Outauais	Avoids the overestimation of the transfer.	Improves the results of trip purpose and destination inference.
Nassir et al. (2011) [17]	ADCS AFC AVL	Geographical and temporal check Transfer time threshold	OD estimation algorithm	Minneapolis-Saint Paul	Relative relaxation of the search in finding the boarding stops.	The transfer time threshold is fixed
Wang et al. (2011) [23]	ADCS AVL	Walking tolerance is 1 km or 12 min.	Trip chaining methodology based on next trip is bus or rail	London	Validates the automatic inference results against large-scale survey results	Linking system usage to home addresses; access behavior could be better understood
Munizaga and Palma (2012) [24] and Munizaga et al. (2014) [25]	AFC GPS AVL	Generalized time	Position-time alighting estimate model	Santiago	(i) Uses generalized time rather than physical distance (ii) Replaces larger on-board survey	The one trip per card destination estimation is missing.
Gordon et al. (2013) [26]	AFC AVL	Walking tolerance is 1 km and max. transfer is 30 min.	Four-step trip chaining algorithm	London	The circuity ratios to decide the potential destination for previous journey.	Not all of the passengers alight from the stops closest to the next journey.
Alsger et al. (2015) [27]	AFC	The dynamic transfer time threshold	OD estimation algorithm	Queensland	Transfer time threshold could be increased.	Extended to compare other estimation methods.

Table 2: Description of each dataset.

Dataset	Description
AFC data	
Card ID	Unique number that could be taken as the passenger ID
O station	Boarding station ID
Entry time	Access time to the station
D station	Alighting station ID
Exittime	Exit time from the station
Station geographical data	
Station ID	Unique station number
Station name	Name of metro station
Station latitude	Latitude of metro station
Station longitude	Longitude of metro station
Station route ID	Route number which serves at metro station
Timetable data	
Service ID	Given number to every trip
Arrival time	Scheduled arrival time
Departure time	Scheduled departure time
Station ID	Given station number
Route ID	Given route number
Ticket fare data	
O station	Entry station ID
D station	Exit station ID
Ticket price	The price for a specific OD pair.

2.2. Data Cleaning and Preparation. It has been highlighted that the level of accuracy of AFC data may vary and the data can be affected by various types of errors. These errors may affect the accuracy of individual journeys and passenger behavior analysis. In the original AFC data, some errors are caused by system failure or passenger error. The data were filtered with some transactions excluded, such as reloaded transactions, transactions with missing information such as no boarding or alighting stops, and transactions with the same entry and exit stations.

As the study uses the trip chaining method to infer the actual destinations and potential purpose, we exclude single trip cardholders due to lack of information. Figure 1 shows the preparation process. With this data process, the destination of every trip leg of each cardholder has been saved in an individual alighting station list which will be used for the trip purpose inference.

3. Methodology

3.1. Assumptions

3.1.1. Trip Purpose for Each Passenger. Trip purpose could be inferred from their alighting and boarding station. For example, if the passenger started his trip at a residential area and went to CBD, we could say that this trip is a work trip. Based on the land use and the daily entry flow pattern for each metro station, we processed the k-means clustering method [29, 30] and classified the stations into 7 categories, and the typical stations are marked in Figure 2.

(1) Working stations (red)

Those stations are usually in the CBD area or near the software plaza. In the morning, commuters take transit to go to work and go back home in the early evening. The morning exit passengers are much larger than that in the afternoon. The entry passenger volume in the early evening or late afternoon is much more than that in the morning. The typical stations such as Guomao Station and Zhongguancun Station are marked in red in Figure 2.

(2) Residential stations (orange)

Beijing has 6 ring roads in the city. The house price is unusually high within the 3rd ring. In order to save living expenses, a lot of citizens go to the 6th or even further place to buy or rent a house. There are some huge residential zones in Beijing such as Huilongguan, Huoying. The passenger flow pattern is the opposite. The morning incoming flow is much larger than that in the afternoon, and most passengers exit at these stations in the afternoon. The typical stations such as Huilongguan Station and Tiantongyuan Station are marked in orange in Figure 2.

(3) Working-residential stations (yellow)

Although the house price is pretty high, comparing with the travel time, some commuters prefer to rent or buy a house in the downtown area. The land use is more like the mix of CBD and the residential place such as the university campus area. The passenger flow patterns of these stations keep stable, and they do not have a flow peak during the day. The typical stations such as Wukesong Station and Gongzhufen Station are marked in yellow in Figure 2.

(4) Transit hub stations (green)

The in-coming and out-coming passenger flows, whether in the morning peak hour or in the afternoon peak hour, are always large in the transit hub. Mostly, they are the key points of the transit line such as transfer stations. The typical stations such as Dongzhimen Station, Xizhimen Station, and Songjiazhuang Station are marked in green in Figure 2.

(5) Railway stations (light blue)

Based on the land use, the railway station is a very independent station category. The in-coming and out-coming flow highly depends on the railway schedule. We have 3 railway stations in Beijing. They are Beijing railway station, Beijing south

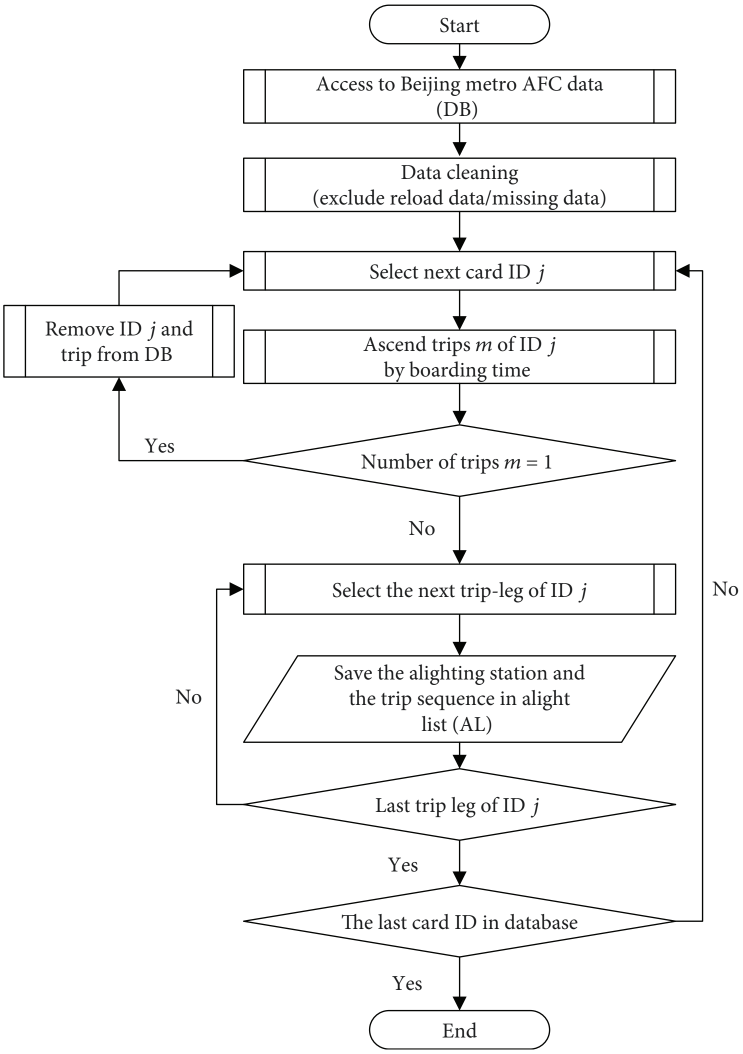

Start
Access to Beijing metro AFC data (DB)
Data cleaning (exclude reload data/missing data)
Select next card ID j
Remove ID j and trip from DB
Ascend trips m of ID j by boarding time
Yes
Number of trips $m = 1$
No
Select the next trip-leg of ID j
No
Save the alighting station and the trip sequence in alight list (AL)
No
Last trip leg of ID j
Yes
The last card ID in database
Yes
End

Figure 1: Data preparation process.

railway station, and Beijing west railway station, which are marked in blue in Figure 2.

(6) Shopping-sightseeing stations (deep blue)

There are some sightseeing and shopping sites such as The Forbidden City and Tiananmen Square, which attract a lot of tourists and visitors every day. For these stations, the total daily passenger volume during the weekends and holidays is usually higher than during workdays. The typical stations for this category, such as Tiananmen East, Tiananmen West, and Xidan stations, are marked in deep blue in Figure 2.

(7) Rural stations (purple)

The Beijing network is a huge network, and the operation distance has reached 608 km. Some rural areas also have operation lines for passengers such as Changing Line and Fangshan Line. The daily average passenger flow is much smaller in the rural lines compared with the volume in the downtown area. The typical rural stations are marked in purple in Figure 2.

For each trip, the trip purpose could be estimated based on the station category. For example, a passenger started his trip from a residential station and finished his trip at a working station. Based on the station category, we could label this trip as a working trip. This process could efficiently determine the trip purpose during the day.

However, there is a category that the station could be a workplace or a residential place. In order to determine the trip purpose for these trips, we performed a filter process. For each passenger in Beijing AFC data, the alighting station and boarding time are recorded according to the alighting station list for a passenger during a week. If the alighting station frequency is more than three times on weekdays, we make an assumption that the passenger is a commuter in the city and this place is a workplace or a home [31]. Considering the trip sequence and boarding time for a serial

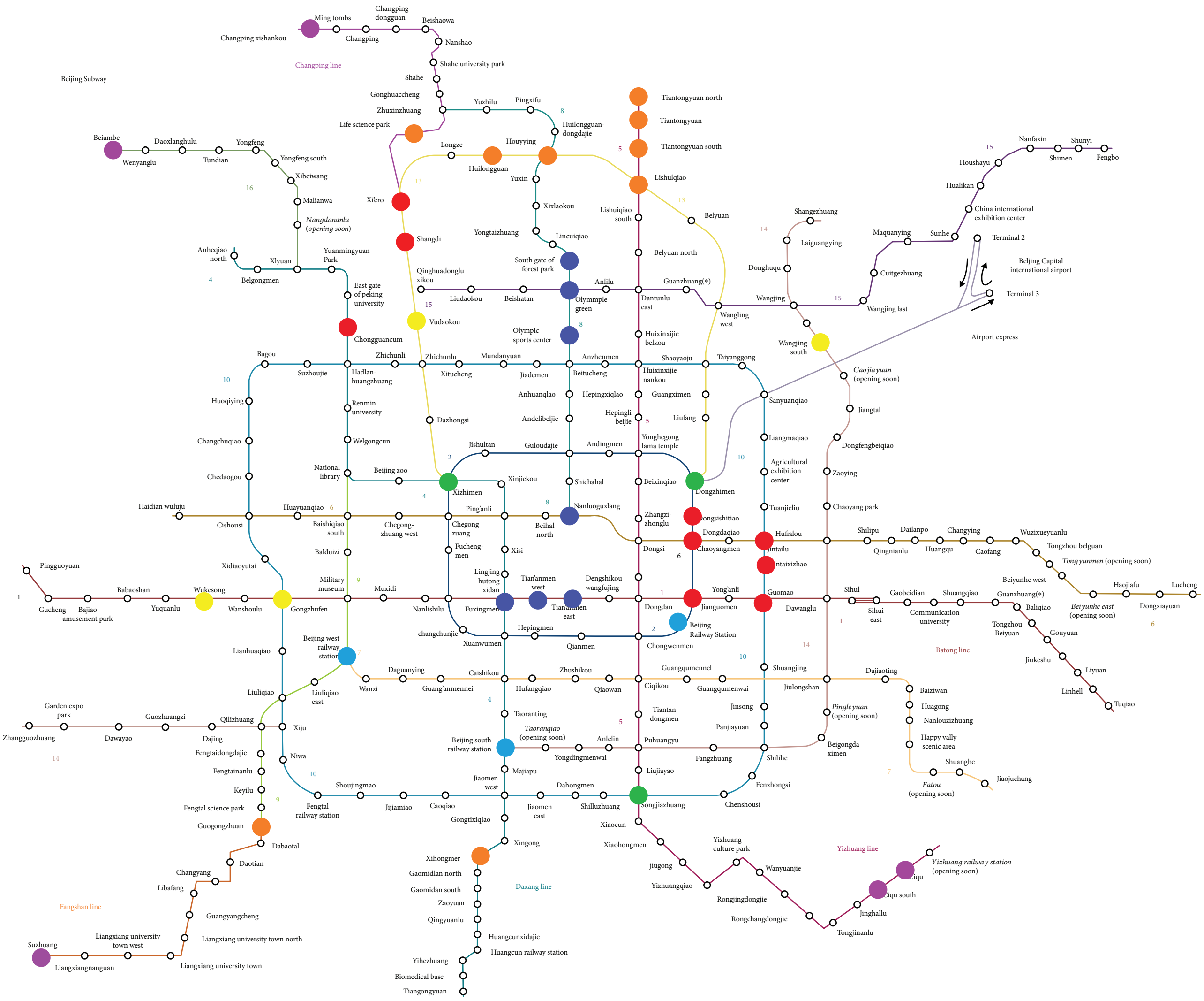


Figure 2: The typical stations for each category in Beijing metro.

number, if the trip happened at the early times of the day and the sequence number is one, we label this trip as a home trip. If the trip occurred later in the day or it is the last trip of the day, we label this trip as a work trip. Figure 3 shows the trip purpose labelling process.

3.1.2. Intelligent Passenger. Although the boarding and alighting information is recorded in the AFC data, the passenger trip routes are not recorded. In our study, we assume every passenger is an intelligent agent and wants to minimize the travel cost and maximize the utility of the travel. As such, the passenger will choose the shortest path from the boarding station to the alighting station. We calculate and use the shortest path travel time as in-vehicle. Also, we assume that a passenger will not detour when they go to another station by foot, so we take the Euclidean distance between the two stations as the walking distance.

3.1.3. The Actual Destination of the Trip and Walking Buffer Circle. AFC data recorded the alighting station, but the actual destination is missing. We assume that the passenger is a smart decision-maker, so he/she would choose an alighting station which is closer to the actual destination. In this study, we assume that the actual destination is somewhere in between the two consecutive stations, the alighting station of the previous trip, and the next trip's origin, as seen in Figure 4(a). However, if the distance between the two consecutive stations is more than a walking threshold (we use 3 km in the later empirical study), shown in Figure 4(b), the passenger is more likely to take other modes of transportation. In this way, the actual destination of the first trip is hard to infer, so we would exclude this trip from the analysis sample.

When the alighting stations are relaxed, in order to find some candidate alighting stations, we set a walking buffer circle. According to the previous literature, we take a 15 min walk, or nearly 1 km, as the walking buffer radius. The stations which are included in the buffer circle are candidate alighting stations, shown as yellow circles in Figure 4(a).

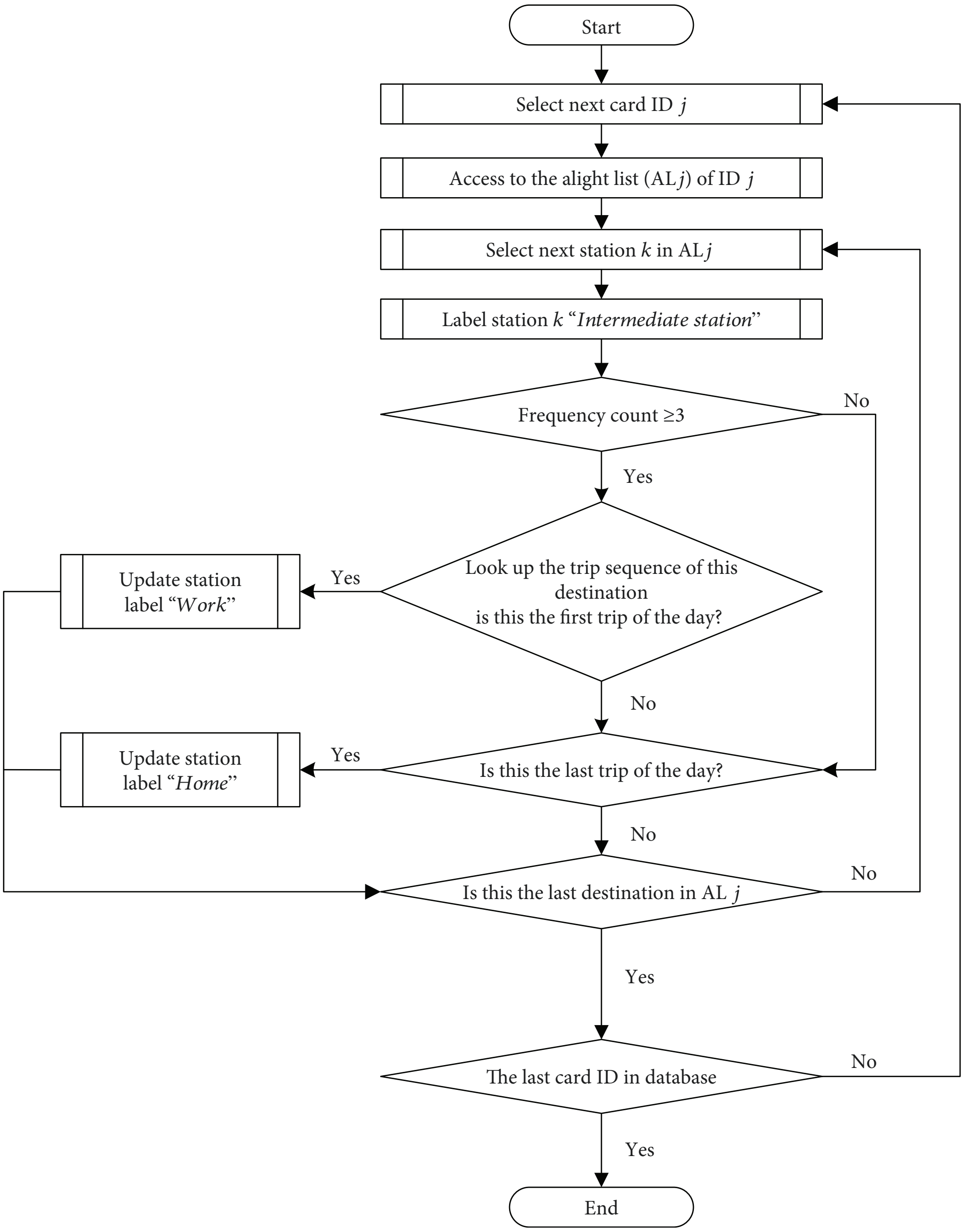

Start
Select next card ID j
Access to the alight list (AL j) of ID j
Select next station k in AL j
Label station k "Intermediate station"
Frequency count ≥3
No
Yes
Look up the trip sequence of this destination is this the first trip of the day?
Yes
Update station label "Work"
No
Is this the last trip of the day?
Yes
Update station label "Home"
No
Is this the last destination in AL j
No
Yes
The last card ID in database
No
Yes
End

Figure 3: Trip purpose labelling process for work-residential trips.

3.2. The Choice Model Specification. The following notation corresponding to the choice model is used:

U_n: Utility function of passenger n

$\alpha_{\mathrm{IVT}}, \alpha_{\mathrm{WT}}, \gamma_{\mathrm{TF}}$: Coefficients for in-vehicle time, walking time and ticket fare, respectively

$T^n_{IVT}, T^n_{WT}, T^n_{TF}$: Value of in-vehicle time, walking time, and ticket for passenger n, respectively

$\eta_{\mathrm{IVT}}, \eta_{\mathrm{WT}}, \eta_{\mathrm{TF}}$: Takes the value one if the corresponding parameter is significant in the utility function

ε: Random error term

J: Choice set for each passenger

T: Factor set

A: Coefficient set

Γ: Trip purpose set. 1, 2, and 3 represent work, home, and others.

3.2.1. Multinomial Logit Model (MNL). The MNL model is the prime model in transportation research which calculated the probability or each choice in a choice set. In Beijing metro, the ticket fare is distance-based, which means that passengers could walk a long distance to save money. When a passenger chooses an alighting station, there are three factors which impact the utility, in-vehicle travel time, walking time, and ticket fare. For each passenger, the utility function can be written as (1), (2), and (3).

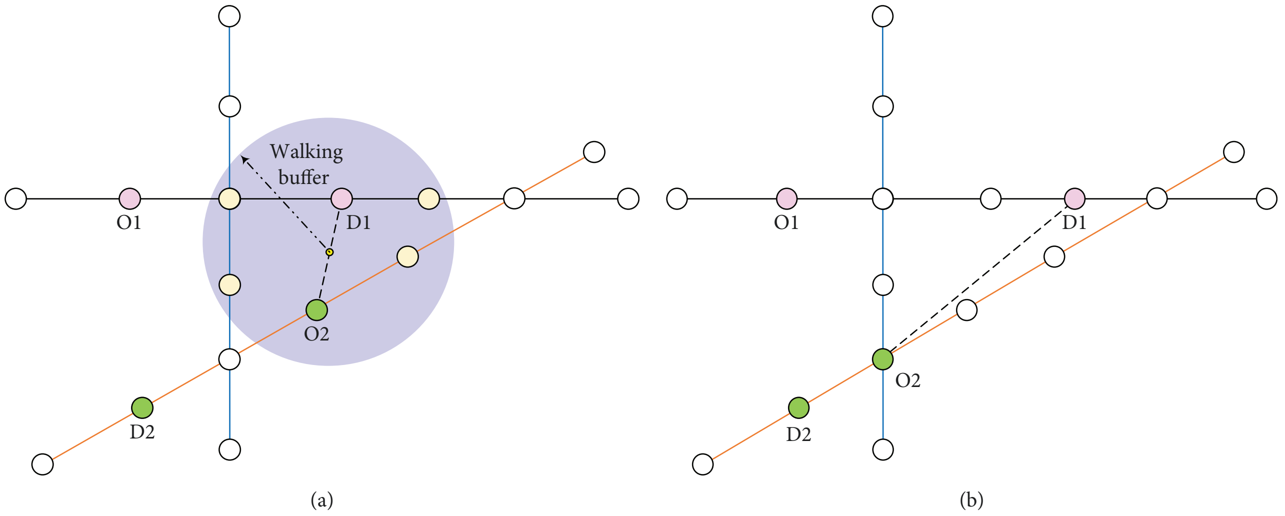


Figure 4: The assumption for actual destination and potential alighting station choices. Pink circles are the boarding and alighting stations of the first trip. Green circles are the boarding and alighting stations of the next trip. Yellow circles are candidate alighting stations.

$$U_n = \alpha_{\mathrm{IVT}}\eta_{\mathrm{IVT}}T^n_{\mathrm{IVT}} + \alpha_{\mathrm{WT}}\eta_{\mathrm{WT}}T^n_{\mathrm{WT}} + \alpha_{\mathrm{TF}}\eta_{\mathrm{TF}}T^n_{\mathrm{TF}} + \varepsilon, \tag{1}$$

$$y_{nj} = \begin{cases} 1 & \text{if } U_{nj} \geq U_{nj'} for j' \in \{1, \dots, J\} \\ 0 & \text{otherwise} \end{cases}, \tag{2}$$

$$\eta_{\mathrm{IVT}},\ \eta_{\mathrm{WT}},\ \eta_{\mathrm{TF}} = \begin{cases} 1 & \text{if} T^n_{\mathrm{IVT}},\ T^n_{\mathrm{WT}},\ \text{and } T^n_{\mathrm{TF}} \text{ are siginificant.} \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

The choice of alternative j by passenger n may be derived from (2) to yield the following functional form of the multinomial logit

$$P\left(y_{nj} = 1 \mid \alpha_{\mathrm{IVT}}, \alpha_{\mathrm{WT}}, \alpha_{\mathrm{TF}}, T\right) = \frac{\exp\left(U_{nj}\right)}{\sum_{j'=1}^{J} \exp\left(U_{nj'}\right)}. \tag{4}$$

3.2.2. Mixed Logit Model with Independent Normally Distributed Random Coefficients. In the standard logit model, the coefficients for the same factors share the same "preference." However, a different passenger could have a different preference for the same factor. Mixed logit models can be derived from a variety of different behavioral specifications, and each derivation provides a particular interpretation. The mixed logit model is defined on the basis of the functional form for its choice probabilities. The utility function in the mixed logit model and the coefficient in (1) are statistical distributions instead of a constant number, which means for each passenger n, α_n, β_n, γ_n follow distributions, and the coefficients vary over people.

$$\alpha_{\mathrm{IVT}n}, \quad \sim f(\alpha_{\mathrm{IVT}} \mid \theta), \quad \alpha_{\mathrm{WT}n} \sim f(\alpha_{\mathrm{WT}} \mid \theta) \quad \alpha_{\mathrm{TF}n} \sim f(\alpha_{\mathrm{TF}} \mid \theta), \tag{5}$$

where θ is the parameter of the distribution over the population, such as the mean and variance of α_n. Conditional on α_n, and assuming the unobserved term ε is iid extreme value, the probability that passenger n chooses alternative j is the standard logit formula.

$$\begin{aligned} L_{nj}(\mathrm{A}_n, T) &= \frac{\exp\left(\alpha_{\mathrm{IVT}}T^{nj}_{\mathrm{IVT}} + \alpha_{\mathrm{WT}}T^{nj}_{\mathrm{WT}} + \alpha_{\mathrm{TF}}T^{nj}_{\mathrm{TF}}\right)}{\sum_{j'=1}^{J} \exp\left(\alpha_{\mathrm{IVT}}T^{nj'}_{\mathrm{IVT}} + \alpha_{\mathrm{WT}}T^{nj'}_{\mathrm{WT}} + \alpha_{\mathrm{TF}}T^{nj'}_{\mathrm{TF}}\right)} \\ &= \frac{\exp\left(\mathrm{A}_{nj}T_{nj}\right)}{\sum_{j'=1}^{J} \exp\left(\mathrm{A}_{nj'}T_{nj'}\right)}. \end{aligned} \tag{6}$$

Different elements in A may follow different distributions (including some being fixed). Because α_n is random and unknown, with the continuous f, the probability should be the integral of the standard logit over the density of A_n.

$$P\left(y_{nj} = 1 \mid \mathrm{A}_n, T\right) = \int L_{nj}(\mathrm{A}, T) f(\mathrm{A} \mid \theta) d\mathrm{A}. \tag{7}$$

3.2.3. Mixed Logit Model with Correlated Normally Distributed Random Coefficients. As in some cases, the different elements in A may be correlated with other elements. For instance, the ticket fare in Beijing metro is distance-based, and the fare distribution could have the correlation with the distribution of in-vehicle time and coming from a joint distribution with respective means and covariance matrix.

$$\sum \alpha_{\mathrm{TF-IVT}} = \begin{bmatrix} \mathrm{var}\ (\alpha_{\mathrm{TF}}) & \mathrm{cov}\ (\alpha_{\mathrm{TF}}, \alpha_{\mathrm{IVT}}) \\ \mathrm{cov}\ (\alpha_{\mathrm{TF}}, \alpha_{\mathrm{IVT}}) & \mathrm{var}\ (\alpha_{\mathrm{IVT}}) \end{bmatrix}. \tag{8}$$

We assume that the in-vehicle time and ticket fare follow a multivariate normal distribution.

$$\begin{pmatrix} \alpha_{\mathrm{TF}} \\ \alpha_{\mathrm{IVT}} \end{pmatrix} \sim \mathrm{MVN}\left(\begin{pmatrix} \overline{\alpha}_{\mathrm{TF}} \\ \overline{\alpha}_{\mathrm{IVT}} \end{pmatrix} \begin{bmatrix} \mathrm{var}\ (\alpha_{\mathrm{TF}}) & \mathrm{cov}\ (\alpha_{\mathrm{TF}}, \alpha_{\mathrm{IVT}}) \\ \mathrm{cov}\ (\alpha_{\mathrm{TF}}, \alpha_{\mathrm{IVT}}) & \mathrm{var}\ (\alpha_{\mathrm{IVT}}) \end{bmatrix}\right). \tag{9}$$

Using the Cholesky factorization [32, 33], the vector $(\alpha_{\mathrm{IVT}}, \quad \alpha_{\mathrm{TF}})^T$ can be replaced by

$$\begin{pmatrix} \alpha_{\mathrm{TF}} \\ \alpha_{\mathrm{IVT}} \end{pmatrix} = \begin{pmatrix} \overline{\alpha}_{\mathrm{TF}} \\ \overline{\alpha}_{\mathrm{IVT}} \end{pmatrix} + \begin{bmatrix} p_{11} & 0 \\ p_{12} & p_{22} \end{bmatrix} \begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix} = \overline{\mathrm{A}} + P\xi, \tag{10}$$

where $\xi_1, \xi_2$are iid standard normal variables and $PP^T = \sum \alpha_{\mathrm{TF-IVT}}$. We applied the three kinds of logit model to the Beijing network to test which one better explains the different perceptions of users.

4. Empirical Study

From the one-week AFC dataset, there were 5.05 million transactions each workday in the Beijing metro system. For a commuter, if he takes the metro to go to work and come back home, he would make at least 2 transactions in the dataset. Averagely, these transactions are made by 2.9 million cardholders, based on the static theory and sample size calculator [34]. The cardholders' sample size is 9573 when the confidence level is 95% and the confidence interval is 1. The sample cardholders made a total of 72,645 trips. For some cardholders, they have some same routine every workday. The repeated routines have the same parameters for each candidate alighting station, so the repeated routine will not affect the coefficient of the logit model. To save the calculation time, in this study, the repeated routines are counted once. After the data cleaning and trip chaining, there are 15,057 distinct trips with inferred destination for the study.

In this study, we choose 1 km as the walking buffer distance [35, 36]. As shown in Figure 3, the candidate alighting stations can be calculated based on the final destination and the location of metro stations. Sometimes, the candidate alighting stations contain more than one category. In order to improve the estimation, we filter the stations by trip purpose. For example, within the walking buffer distance, there are 5 candidate alighting stations from A to E. We already know that this trip is a work trip. If 5 stations all belong to working stations, the five stations are all candidate alighting stations. If station B is a home station, we will keep the other 4 stations as the candidate alighting stations.

For some OD pairs, the distance between real alighting station and alternative alighting station is more than 1 km, and these OD pairs did not have candidate alighting stations, which means the passenger could only egress at that station. The logit model could not be estimated in these no-candidate alighting stations or only one alighting station case. Therefore, these records are excluded, after which 13,180 trips remained.

After applying the trip purpose labelling process, 6027 trips are labeled as work trips, 2339 trips are home trips, and the remaining 4814 trips have other purposes. We used Biogeme [37] to estimate the model coefficients.

4.1. MNL Results. For the utility function, we made the assumption that the passenger choice may be influenced by in-vehicle time, walking distance, and ticket fare. To make sure which of these factors significantly impact the utility, we tried every factor and their combination in the model to determine which ones are mostly considered in the choice process. Table 3 excludes the results with a p value over 0.05 and shows the results of the combination of different factors for the different trip purposes.

Firstly, we consider the only single impact factor in the utility function. We found out that a single factor could not explain the passenger behavior very well, especially for the ticket fare, which did not influence the passenger choice. The walking time is more influential among three factors. The coefficient for in-vehicle time is almost the same for four types of the trips, but the coefficient for walking time differs based on different trip purposes.

For the two-factor combinations, in-vehicle time and walking time explained the user behavior as the best among the three possible combinations. This combination could illustrate every trip purpose well. Regardless of the trip purpose, there is higher disutility associated with walking time compared with in-vehicle time. On average, the walking and in-vehicle time coefficient ratio $\alpha_{\mathrm{WT}}/\alpha_{\mathrm{IVT}}$ is 1.462. However, the sensitivity for walking time is different based on the trip purpose. Work trips have the highest penalty for walking, and the coefficient ratio is 1.635 while the coefficient ratio for home trips and other trips is 1.212 and 1.149, respectively.

As for the final log likelihood, the chi-square test was used to analyze the passenger behavior based on different trip purposes rather than overall. In this case, we use $\alpha = 0.05$ as the confidence interval. After checking the χ^2 distribution table, $\chi^2_{0.05,3} = 7.815$, compared with $|\sum_{i=1}^{3} \mathrm{FLL} - \mathrm{FLL}_{\mathrm{total}}| = 55.14 > 7.815$, which indicates it is more appropriate to analyze the passenger behavior based on different trip purposes rather than overall analysis.

When we only consider the rho square, the model which has three factors in the utility function performs a little better than the two-factor combinations. But in the three-factor combination model, the coefficient for ticket fare is positive. In the Beijing metro system, the ticket fare is distance-based with a potentially high correlation with in-vehicle time. So, we could consider the positive coefficient as an adjustment for overestimation of the in-vehicle time coefficient. To be more objective, in the next step, the walking and in-vehicle time model will be as the test model for home, work, other, and total trips, and the three-factor model will be the candidate model for work, other, and total trips.

4.2. Mixed Logit Model Results. We considered the three-factor and two-factor models in the mixed logit model for utility function estimation. For each utility function, similar to the MNL analysis, we test the factors with different combinations such as single-factor or two-factor with independent or correlated distributions.

4.2.1. Three Factors in Utility Function. In-vehicle time, walking time, and ticket fare are all considered in the three-factor utility function. For each trip purpose, fourteen combinations of the mixed logit model were tested. Because of the

Table 3: Results of the different factor combination of the MNL model.

	Purpose	RhS	ILL	FLL	TF_Coff		IVT_Coff		WT_Coff	
					MV	PV	MV	PV	MV	PV
Single factor IVT	W	0.065	−7213.23	−6812.13	—	—	−7.37	0.00	—	—
	H	0.065	−2847.97	−2670.73	—	—	−7.12	0.00	—	—
	O	0.064	−6065.24	−5725.43	—	—	−7.74	0.00	—	—
	T	0.065	−16071.30	−15277.80	—	—	−7.52	0.00	—	—
Single factor WT	W	0.412	−7213.23	−4466.97	—	—	—	—	−18.2	0.00
	H	0.131	−2847.97	−2472.62	—	—	—	—	−9.87	0.00
	O	0.283	−6065.24	−4435.62	—	—	—	—	−13.2	0.00
	T	0.376	−16071.30	−10801.90	—	—	—	—	−16.4	0.00
Two factors WT and IVT	W	0.475	−7213.23	−4015.23	—	—	−11.2	0.00	−18.3	0.00
	H	0.21	−2847.97	−2442.40	—	—	−7.4	0.00	−9.01	0.00
	O	0.353	−6065.24	−3979.53	—	—	−13.2	0.00	−15.2	0.00
	T	0.414	−16071.30	−9729.70	—	—	−11.2	0.00	−16.4	0.00
Three factors	W	0.477	−7213.23	−4008.45	0.570	0.02	−11.4	0.00	−19.8	0.00
	O	0.354	−6065.24	−3960.00	0.665	0.00	−13.7	0.00	−15.3	0.00
	T	0.412	−16071.30	−9705.12	0.598	0.00	−13.1	0.00	−16.6	0.00

RhS = rho square; ILL = init log likelihood; FLL = final log likelihood; PV = p value; MV = mean value; W = work purpose; H = home purpose; O = other purpose; T = total trip, did not distinguish trip purpose.

computational complexity of mixed logit model estimation, only some cases could reach convergence, such as the two independent distributions for fare and in-vehicle time. However, for some combinations, even when the estimation is converged, the coefficients in the model did not pass the p value test so the model did not provide a good interpretation of the passenger behavior. Based on the convergence and p value test, only two models passed. The first one is the single walking time distribution model, which explained every trip purpose except home trips. The second one is a two-independent distribution (walking time and ticket fare) model, which only explains the total sample. No model among fourteen combinations passed for home purpose trips.

Among the passed models, the penalty for walking time is much higher than that for in-vehicle time, where the home trip has the highest coefficient ratio. Meanwhile, from other mixed logit models, we learned that the ticket fare standard deviation and in-vehicle time standard deviation are not significant for the utility function, which means that different passengers could share the same coefficient for ticket fare and in-vehicle time.

4.2.2. Two Factors in Utility Function. From the previous tests, we learned that walking time and in-vehicle time are more important factors compared with ticket fare. In this case, we only consider the walking and in-vehicle times in the utility function to see which mixed logit combinations could explain the passenger behavior well. From the results, similar to the three-factor utility condition, the single walking time distribution model also passed the p value and convergence test this time, which also explained every trip purpose except home trips. The second passed model is the independent distribution combination for walking time and in-vehicle time, which performed well for other trip purposes.

Above all, for the work trips, other trips, and total trips, some mixed logit models could illustrate passenger behavior well and based on the rho square, mixed logit models performed a better estimation result than MNL models did. Comparing the models with rho square and p value for each coefficient, the three-factor models performed better than the two-factor models did. The selected mixed logit model for alighting station choice estimation is shown in Table 4. The single walking time distribution utility function, which is a three-factor model, is selected for work trips and other purpose trips, and the two-independent distribution (walking time and ticket fare) utility function which is a three-factor mixed logit model is selected for the total trip estimation.

4.3. Alighting Station Estimation. According to the research above, we selected the best model that could illustrate every trip purpose. This time, we randomly select another 9573 cardholders and did the same prework such as data cleaning, trip purpose labelling, and candidate station selection as presented in the first part of the empirical study. For each trip purpose, 70% of the data is used as the sample to estimate the coefficient for each model and the remaining data is used for alighting station estimation simulation by Biosim [37]. The percentage of records for which the alighting station could be estimated correctly compared with the AFC records is shown in Table 5.

From Table 5, in general, regardless of the trip purpose, approximately 71.9% of the alighting stations could be estimated correctly by the MNL model and approximately 78.6% by the mixed logit model, which performed better when estimating the alighting stations. For the different trip

Table 4: The selected combination of mixed models for different trip purposes.

Pur	Mixed logit model	RhS	ILL	FLL	TF_Cofficient		IVT_Cofficient		WT_Cofficient		TF_Stad		WT_Stad	
					MV	PV	MV	PV	MV	PV	MV	PV	MV	PV
W	three-factor utility function, WT distribution	0.627	−7213.23	−3202.92	0.81	0.08	−18.1	0.00	−249.00	0.00	—	—	−156.00	0.02
O	Three-factor utility function, WT distribution	0.512	−6065.24	−3212.77	1.32	0.02	−21.1	0.00	−312.00	0.08	—	—	−266.00	0.08
T	Three-factor utility function, independent distributions WT and TF	0.593	−16071.3	−8010.65	1.03	0.00	−17.3	0.00	−302.00	0.00	5.62	0.04	−213.00	0.00

Stad = standard deviation.

Table 5: Results for alighting station estimation based on selected MNL and mixed logit models.

Trip purpose	MNL		Mixed logit	
	Model	Percentage	Model	Percentage
Home	Two factors (WT and IVT)	66.30%	—	—
Work	Two factors (WT and IVT)	78.27%	Three-factor utility function, WT distribution	81.31%
Others	Two factors (WT and IVT)	70.74%		75.35%
Total	Two factors (WT and IVT)	72.59%	Three-factor utility function, independent distributions WT and TF	79.23%

purposes, the simulation for home trips did not perform very well and only 66.30% of the alighting stations were estimated correctly. When we map the errors on the Beijing network, we found out that the incorrect estimations are mostly around the big residential zones which are surrounded by a lot of metro stations. Because of the low penalty for walking for home trips, the alighting stations for home trips could be more flexible. This potentially could affect the estimation results for home trips. For the work trips, the MNL logit model and mixed logit model both worked best among other trip purpose simulations, likely because work trips are more predictable due to their regular patterns. For the other trips, the mixed logit model performed better than the MNL model did because the mixed logit model could illustrate passengers' deviation more properly than the MNL model could.

5. Conclusions

This study is focused on the utility function calibration for alighting station estimation for different trip purposes. The main conclusions of this paper are fivefold:

(1) We provided a two-step trip purpose labelling process to infer the trip purpose. Based on the land use and passenger flow pattern, *k*-means clustering was applied to classify the stations into 7 categories. For the working-residential stations, we use the trip time and alighting station frequency to infer the trip purpose.

(2) The walking buffer radius was applied to infer the real destination. With three assumptions and the trip chaining method, the actual destination and candidate alighting stations of the trips were inferred.

(3) The MNL mixed logit models were proposed to illustrate passenger behavior. In order to estimate alighting stations, MNL and mixed logit models with different combinations of independent variables were discussed to illustrate passenger behavior for different trip purposes.

(4) The influence factors for alighting station choice were tested. In the empirical study, passengers were found to have a different penalty for walking time and in-vehicle time based on trip purpose, and in general, walking time has a higher disutility. Ticket fare was not found significant compared with walking time and in-vehicle time.

(5) The validation test represents the feasibility of the methodology proposed in this paper. Using a validation test, the model could successfully estimate 75% of the alighting stations. The work purpose trips have higher accuracy compared with other purpose trips. This coefficient calibration helps planners understand passenger behavior better and could be used in planning and policy applications.

This research, with the real AFC alighting station data, provided a new method to infer the alighting station and could validate the passenger behavior. Comparing with the on-board survey, this one is much cheaper and more convenient. Meanwhile, this work considers the passenger alighting behavior with different trip purposes, which is a new aspect of alighting behavior analysis.

Some aspects of this study could be improved in future research. The trip purpose labelling process is based on land use, passenger flow pattern, trip time, and alighting station frequency. We can define the trip purpose as a latent variable and apply the latent logit model to capture the trip purpose based on alighting station frequency, trip sequence, and boarding time automatically. Moreover, we will apply the model to a bigger data sample in order to make a more accurate estimation of complex models such as mixed logit. Finally, if possible, passengers' sociodemographic characteristics could be incorporated in the choice model to make the choice more interesting and analyze passenger behavior in a different way.

Conflicts of Interest

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This work was supported by the Foundation of China Scholarship Council, the National Natural Science Foundation Project of P.R. China under Grant no. U1434207, and the Beijing Municipal Natural Science Foundation under Reference no. 8162033. The authors thank the coworkers in CEGE transit lab at University of Minnesota for their

constructive suggestions and comments that have led to a significant improvement in this paper.

References

[1] "Wikipedia List of smart card," https://en.wikipedia.org/wiki/List_of_smart_cards.

[2] 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.

[3] J. A. Petsinger, "Electromagnetic shield to prevent surreptitious access to contactless smartcards," US Patent 6,121,544, 2000.

[4] Smart Card Alliance, *RF-Enabled Applications and Technology: Comparing and Contrasting RFID and RF-Enabled Smart Cards*, Smart Card Alliance Identity Council, 2007.

[5] 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.

[6] J. Y. Park, D.-J. Kim, and Y. Lim, "Use of smart card data to define public transit use in Seoul, South Korea," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2063, no. 1, pp. 3–9, 2008.

[7] M. Trépanier and C. Morency, "Assessing transit loyalty with smart card data," in *12th World Conference on Transport Research*, pp. 11–15, Lisbon, Portugal, July 2010.

[8] P. White, M. Bagchi, H. Bataille, and S. M. East, "The role of smartcard data in public transport," in *Proceedings of the 12th World Conference on Transport Research*, pp. 1–16, Lisbon, Portugal, 2010.

[9] B. Agard, C. Morency, and M. Trépanier, "Mining public transport user behaviour from smart card data," *IFAC Proceedings Volumes*, vol. 39, no. 3, pp. 399–404, 2006.

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

[11] M. Hofmann, S. P. Wilson, and P. White, "Automated identification of linked trips at trip level using electronic fare collection data," in *Transportation Research Board 88th Annual Meeting. No. 09-2417*, pp. 1–18, Transportation Research Board Meeting, 2009.

[12] 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, pp. 119–126, 2006.

[13] M. Trépanier, C. Morency, and C. Blanchette, "Enhancing household travel surveys using smart card data?," in *88th Annual Meeting of the Transportation Research Board, Washington*, pp. 85–96, Transportation Research Board Meeting, 2009.

[14] 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.

[15] M. Trepanier and F. Vassiviere, "Democratized smartcard data for transit operator," in *15th World Congress on Intelligent Transport Systems and ITS America's 2008 Annual Meeting ITS America ERTICOITS Japan Trans Core*, pp. 1838–1849, World Congress on Intelligent Transport Systems and its Americas meeting, 2008.

[16] Y. Sun, M. Hrušovský, C. Zhang, and M. Lang, "A time-dependent fuzzy programming approach for the green multimodal routing problem with rail service capacity uncertainty and road traffic congestion," *Complexity*, vol. 2018, Article ID 8645793, 22 pages, 2018.

[17] 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.

[18] 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, pp. 183–187, 2002.

[19] 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.

[20] J. Zhao, *The Planning and Analysis Implications of Automated Data Collection Systems: Rail Transit OD Matrix Inference and Path Choice Modeling Examples, [Ph.D. Thesis]*, Massachusetts Institute of Technology, 2004.

[21] 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.

[22] K. K. A. Chu and R. Chapleau, "Enriching archived smart card transaction data for transit demand modeling," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2063, no. 1, pp. 63–72, 2008.

[23] 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.

[24] 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.

[25] M. Munizaga, F. Devillaine, C. Navarrete, and D. Silva, "Validating travel behavior estimated from smartcard data," *Transportation Research Part C: Emerging Technologies*, vol. 44, pp. 70–79, 2014.

[26] J. B. Gordon, H. N. Koutsopoulos, N. H. M. Wilson, and J. P. Attanucci, "Automated inference of linked transit journeys in London using fare-transaction and vehicle location data," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2343, no. 1, pp. 17–24, 2013.

[27] A. A. Alsger, M. Mesbah, L. Ferreira, and H. Safi, "Use of smart card fare data to estimate public transport origin–destination matrix," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2535, pp. 88–96, 2015.

[28] A. Tavassoli, A. Alsger, M. Hickman, and M. Mesbah, "How close the models are to the reality? Comparison of transit origin-destination estimates with automatic fare collection data," in *Australasian Transport Research Forum 2016 Proceedings*, pp. 1–15, Melbourne, VIC, Australia, 2016.

[29] B. Gao, Y. Qin, X. M. Xiao, and L. X. Zhu, "K-means clustering analysis of key nodes and edges in Beijing subway network," *Journal of Transportation Systems Engineering and Information Technology*, vol. 14, no. 3, pp. 207–213, 2014.

[30] Z. Yue, F. Chen, Z. Wang, J. Huang, and B. Wang, “Classifications of metro stations by clustering smart card data using the Gaussian mixture model,” *Urban Rapid Rail Transit*, vol. 107, no. 2, pp. 48–51, 2017.

[31] S. Shizhao, *Operation Performance Assessment of Urban Rail Transit Based on Travel Time Delay*, Beijing Jiaotong University, Beijing, China, 2016.

[32] C. N. Haddad, *Cholesky factorization*, Springer, 2001.

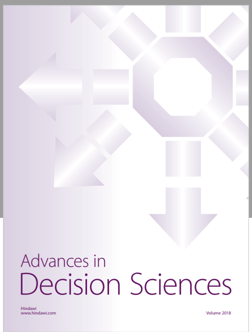
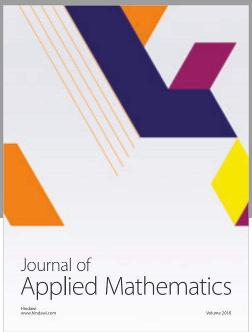
[33] R. B. Schnabel and E. Eskow, “A new modified Cholesky factorization,” *SIAM Journal on Scientific and Statistical Computing*, vol. 11, no. 6, pp. 1136–1158, 1990.

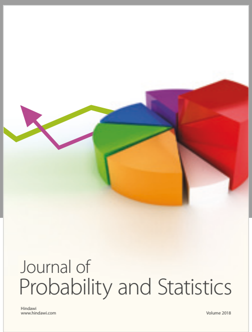
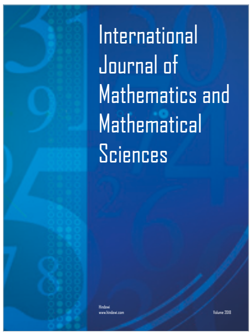
[34] “Sample Size Calculator,” http://www.calculator.net/sample-size-calculator.html.

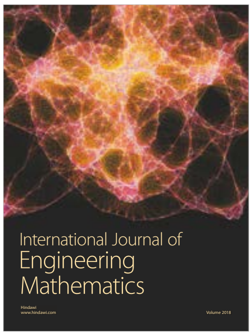
[35] S. O’Sullivan and J. Morrall, “Walking distances to and from light-rail transit stations,” *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1538, pp. 19–26, 1996.

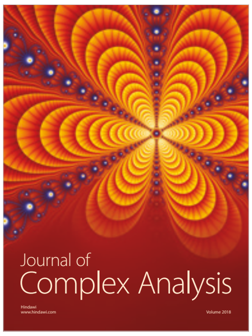
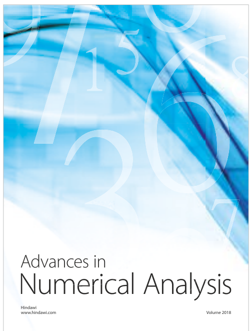
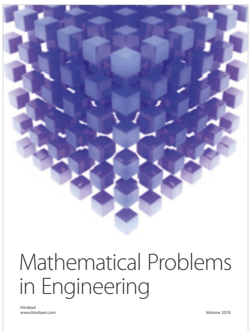
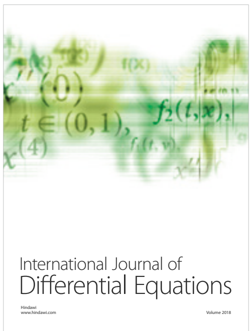
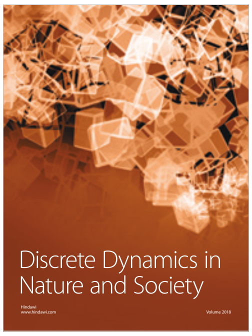
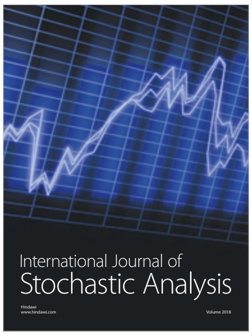
[36] K. Lu, B. Han, and X. Zhou, “Smart urban transit systems: from integrated framework to interdisciplinary perspective,” *Urban Rail Transit*, vol. 4, no. 2, pp. 49–67, 2018.

[37] “Biogeme and Biosim,” http://biogeme.epfl.ch/.

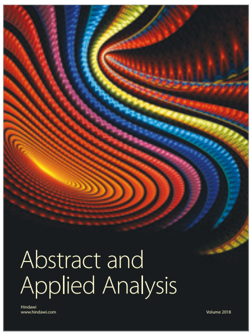
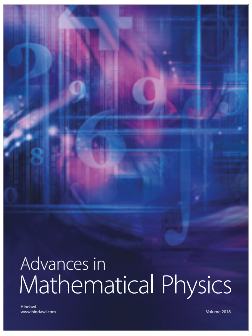
Advances in
Operations Research

Advances in
Decision Sciences

Journal of
Applied Mathematics

The Scientific
World Journal

Journal of
Probability and Statistics

International
Journal of
Mathematics and
Mathematical
Sciences

Journal of
Optimization

International Journal of
Engineering
Mathematics

International Journal of
Analysis

Journal of
Complex Analysis

Advances in
Numerical Analysis

Mathematical Problems
in Engineering

International Journal of
Differential Equations

Discrete Dynamics in
Nature and Society

International Journal of
Stochastic Analysis

Journal of
Mathematics

Journal of
Function Spaces

Abstract and
Applied Analysis

Advances in
Mathematical Physics