Hindawi
Journal of Advanced Transportation
Volume 2019, Article ID 6830450, 15 pages
https://doi.org/10.1155/2019/6830450

Research Article

Data-Driven Approaches to Mining Passenger Travel Patterns: "Left-Behinds" in a Congested Urban Rail Transit Network

Xing Chen,[1] Leishan Zhou,[1] Zixi Bai,[1] Yixiang Yue,[1] Bin Guo,[2] and Hanxiao Zhou[1]

[1]*Department of Transportation Management Engineering, School of Traffic and Transportation, Beijing Jiaotong University, China*
[2]*National Research Center of Rail Transit and Transportation Training and Accreditation, Beijing Jiaotong University, China*

Correspondence should be addressed to Zixi Bai; 11114230@bjtu.edu.cn

Received 20 November 2018; Accepted 6 March 2019; Published 1 April 2019

Academic Editor: Eneko Osaba

The "left-behind" phenomenon occurs frequently in Urban Rail Transit (URT) networks with booming travel demand, especially during peak hours in a complex URT network, which makes passenger travel patterns more complicated. This paper proposes a methodology to mine passenger travel patterns based on fare transaction records from automatic fare collection (AFC) systems and Automatic Vehicle Location (AVL) data from Communication Based Train Control (CBTC) Systems or tracking systems. By introducing the concept of a sequence, a space-time-sequence trajectory model is proposed to simulate a passenger's travel activities, including when they are left-behind. The paper analyzes passenger travel trajectory links and estimates the weight of each feasible trajectory under tap-in/tap-out constraints. The station time parameters, including access/egress and transfer walking-time parameters, are important inputs for the model. The paper also presents a maximum-likelihood approach to estimate these parameters from AFC transaction data and AVL data. The methodology is applied to a case study using AFC and AVL data from the Beijing URT network during peak hours to test the proposed model and algorithm. The estimation results are consistent with the results obtained from the authorities, and this finding verifies the feasibility of our approach.

1. Introduction

During the last decade, Urban Rail Transit (URT) in Mainland China has developed from a total system length of only 763 kilometers ten years ago to 5033 kilometers by the end of 2017 [1]. With the rapid development of the URT network, travel demand has also experienced a booming increase. In the past 10 years, the average daily passenger traffic of the Beijing URT system has increased from 1.92 million in 2007 to 10.35 million in 2017, an increase of 439% [2]. The Mass Transit Railway system (MTR) in Hong Kong has increased approximately 131.6% in patronage since 2006 [3].

The significant increase in travel demand has resulted in congestion and overcrowding both in stations and in train vehicles; this has become a serious problem for URT operators to address, particularly during peak hours. On one hand, congestion brings security risks. On the other hand, congestion and overcrowding reduce the attractiveness of the URT network, and some passengers will choose other modes of transportation. Additionally, a new phenomenon appears that we term "left-behind"; some passengers fail to board the first departing train after their arrival at a platform and must wait for a later one. This occurs mainly because the travel demand exceeds the network supply during a given operational time interval due to train vehicle capacity.

To address the above problems, numerous methods have been proposed and adopted from both the operator's perspective and the passenger's perspective. Guan (2013) [4] and Yu et al. (2015) [5] developed a model for network design with the objective of minimizing the number of transfers. Niu and Zhou [6] presented a methodology to optimize the timetable to minimize waiting time under time-dependent travel demands and oversaturation. They analyzed the characteristics of passenger flow and formulated a model to minimize passengers' waiting times or minimize the number of transfers. To better grasp the distribution of URT network passenger flow, methodologies to study passengers' travel patterns have been developed. These facilitate a number of applications, including (i) analysis of passengers' path-choice preferences, such as minimum time and minimum number

of transfers, (ii) prediction of individual passengers' locations and the future distribution of URT network passenger flow, (iii) optimizing train scheduling both from the subway line level by identifying the most congested stations and sections and from the network level by identifying the transfer hot-spots, and (iv) guiding passengers to avoid congested sections as much as possible by informing route suggestions, congestion levels, etc.

Thus, this paper attempts to mine passenger travel patterns based on automatic fare collection (AFC) transaction data and Automatic Vehicle Location (AVL) data. It builds on our prior work on the problem [7] and reconstructs a passenger's trajectory by introducing the concept of a sequence to describe left-behind. The prior work ignored train vehicle capacity constraints and assumed that all passengers left a platform for another subway station by the first departing train after their arrival at the platform. We propose the concept of sequence to describe the relationships between passenger's arrival and departure of trains. By separating time periods into segments according to stations, train directions, and the departure times of trains passing stations, this paper reconstructs a passenger's trajectory and proposes a space-time-sequence trajectory model. The model generates a set of feasible space-time-sequence trajectories that indicate a passenger's precise travel patterns and the expected number of times a passenger is left-behind on a platform. Then, a methodology is presented to estimate the number of left-behinds and the probability of each trajectory based on the distribution of station access/egress walking time and transfer walking time. These distributions can be obtained from manual surveys. However, these require substantial labor. The paper presents a maximum-likelihood estimation methodology based on passengers' trajectories.

The main contributions of this study include the following:

(i) A space-time-sequence trajectory model to simulate a passenger's travel itinerary in a congested URT system. The model introduces the concept of sequence and provides a means to indicate the left-behind phenomenon.

(ii) A maximum-likelihood estimation methodology based on AFC and AVL data that estimates passenger's travel patterns. More automatic data instead of empirical data or manual survey data is used in the methodology; this minimizes the occasional deviation caused by human factors. Additionally, it reduces the difficulty of obtaining data.

(iii) A data-driven method to estimate station walking-time parameters and the expected number of times passengers are left-behind. Station walking-time parameters, including access walking time and egress walking time, are the basic parameters for a station. They are usually obtained by manual survey or observation; however, these require excessive labor and cost. The method proposed in this paper estimates these parameters using statistical methods based on passengers' space-time-sequence trajectories.

The remainder of this study is organized as follows. Section 2 reviews relevant studies in the literature. Section 3 introduces the main idea of mining passengers' travel patterns and illustrates an example. Section 4 describes the model, followed by the introduction of the solution algorithm in Section 5. Section 6 presents a numerical experiment on a real-world network. The final section provides our conclusions and suggestions for future research directions.

2. Literature Review

Many scholars and researchers have studied URT network passenger travel patterns during the last decades. At the beginning of these studies, it was generally not possible to obtain bulk data, including passengers' tap-in/tap-out information and actual train movement data. Because of the lack of data, numerous methods were proposed at macroscale level. These methodologies mainly analyzed passengers' travel patterns from a network-flow perspective. They generated sets of feasible paths for each origin-destination (OD) and assigned passenger flows to each path following specific principles. The three most well-known principles are the all-or-nothing principle, the stochastic-assignment principle [8–12], and the user-equilibrium-assignment principle [13–18].

With the wide adoption of AFC systems and rapid development of train tracking systems such as the Communication Based Train Control (CBTC) System, massive detailed data was collected and saved to databases. Most AFC systems record passenger tap-in/tap-out information accurately except for some cases such as the New York City Transit Authority (NYCT) system, in which exit swipe information is not recorded. The Automatic Vehicle Location (AVL) system records train arrival and departure times at stations accurately and in detail. With bulk data collected daily automatically and continuously, some novel methodologies for transit performance [19–21] and management [6, 22, 23] have been developed.

Dai (2015) [22] presented a multimodal evacuation model for metro disruptions based on AFC data in Shanghai, China. Using AFC data of stations in urban areas of Hong Kong, Wang (2015) [21] developed a methodology to analyze metro trip patterns at an aggregate level. Kusakabe and Asakura (2014) [24] estimated passengers' behavioral attributes of trips with a data fusion methodology using smart cards. They observed and compared continuous long-term changes in passengers' trips as well as personal trip survey data and constructed a Bayes probabilistic model to estimate the purposes of passengers' trips. Jin (2015) [19] evaluated transit service performance by developing a data-mining logic methodology based on transit smart-card data.

Some scholars and researchers have analyzed network passenger flow at the individual level. Some have proposed a specific trajectory model to simulate a passenger's trip activities and estimated the maximum-likelihood path based on tap-in/tap-out constraints; numerous assumptions are embedded into stages of the models' building process. Poon M.H. (2004) [25] assumed that all passengers have full predictive information about present and future network conditions. Chen (2018) [7] assumed that all passengers can always board the first train arriving. Some additional input parameters are needed and have a significant impact on the accuracy of the estimation result.

Sun (2016) [26] presented a schedule-based passenger's path-choice estimation model for a multioperator rail transit network using automatic fare collection data. In this paper, a Train Schedule Connection Network (TSCN) was constructed, and the estimated passenger path-choice was converted to the problem of generating a feasible set of network paths. The Fail-to-Board (FtB) phenomenon was modeled and the weight of each path could be calculated based on the set of feasible paths. The accuracy of the result was highly dependent on inputs, such as FtB parameters. However, these parameters cannot be obtained directly and are not easy to calculate.

Poon M. H. et al. (2004) [25] present a schedule-based transit model to solve passenger assignments for a congested network. They assumed that all passengers have full predictive information about present and future network conditions and always travel by the minimum-cost path. However, it is not possible for passengers to be informed of full information about network conditions. Furthermore, the minimum-cost path is time-dependent. Frequent URT passengers have their respective perceptions for choosing paths, gaining experience day by day. Additionally, the tap-out times are not taken into account when loading network flow.

Timon Stasko (2015) [20] analyzed passengers' ridership at the train level using actual train movement data. He built a customized network representation by estimating train movements and developing an origin-destination table. The methodology formulates a trip trajectory with 10 types of arcs. It assigns passengers to trains using a Frank-Wolfe approach, with customizations designed for transit. However, it is unnecessary to infer a destination for most of the URT network. Finally, the accuracy of the result is highly dependent on boarding penalties.

3. Problem Description

Time and space are two important attributes of passenger travel. Although AFC systems record passenger transaction information in detail, including precise transaction times and locations when the passenger swipes his/her smart card, detailed information on a passenger's itinerary is not included. This section describes methods for estimating a passenger's detailed travel information in both time and space.

Space-time models attempt to integrate travellers' time-dependent movements/trajectories with the transportation network and are widely used in transportation geography modeling literature. A space-time trajectory indicates a passenger's movements among activity locations with respect to time, providing a useful means to describe both the spatial and temporal aspects of a passenger's travel status. However, the key focus of this paper, the number of left-behinds due to crowding, cannot be obtained directly from a specific space-time trajectory. To estimate the number of times passengers are left-behind, a parameter defined as a sequence is introduced, and a passenger space-time-sequence trajectory model is developed.

The time duration of the study is segmented into a set of successive intervals based on trains' departure times, stations, and directions, such as upward and downward directions. For example, there are a total of 55 trains passing a station in the upward direction between 7:00 AM and 9:00 AM, and the first/last train's departure time is 07:02:30/08:59:00. Thus, the time period is segmented into 56 successive intervals in the upward direction; the sequence number of the first time interval from 07:00:00 to 07:02:30 should be 1. The sequence number of the last time interval from 08:59:00 to 09:00:00 should be 56. The sequence numbers of these intervals should also be successive and the sequence number of any time interval should be smaller than that of any later time interval.

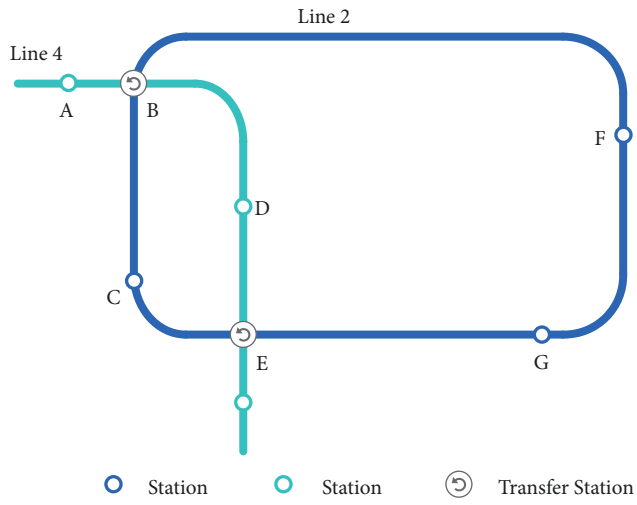

Figure 1: URT network topology.

Assume that there is a passenger travelling from Station A to Station G through the URT system shown in Figure 1 during peak hours.

Obviously, there are three feasible spatial routes for this trip: (i) A ⟶ B ⟶ C ⟶ E ⟶ G; (ii) A ⟶ B ⟶ F ⟶ G; (iii) A ⟶ B ⟶ D ⟶ E ⟶ G. The passenger transfers at Station B if travelling by route (i) or (ii), whereas he/she transfers at Station E if travelling by route (iii). For this journey, the passenger needs to experience the following activities: (i) moving from the entry gate to the platform, (ii) waiting at platforms of both the origin and transfer stations, (iii) transfer from Line 4 to Line 2, (iv) egress from platform to exit gate, and (v) on-train. The passenger may be left-behind and unable to board the first-arriving train to leave the platform at both the origin station and the transfer station during peak periods because of crowding on the platforms and in the train vehicles. Figure 2 illustrates the theoretically feasible space-time-sequence trajectories from Station A to Station G under tap-in/tap-out constraints.

As shown in Figure 2, the concept of an ideal boarding node is proposed to distinguish platform waiting from left-behind. An ideal boarding node represents a passenger's theoretically earliest boarding activity; i.e., it represents the passenger boarding the first departure train after his/her arrival at a platform. The arc linking the entry space-time-sequence node and ideal boarding node indicates that a passenger moves from the entry gate to the platform and waits until the first train leaves after his/her arrival.

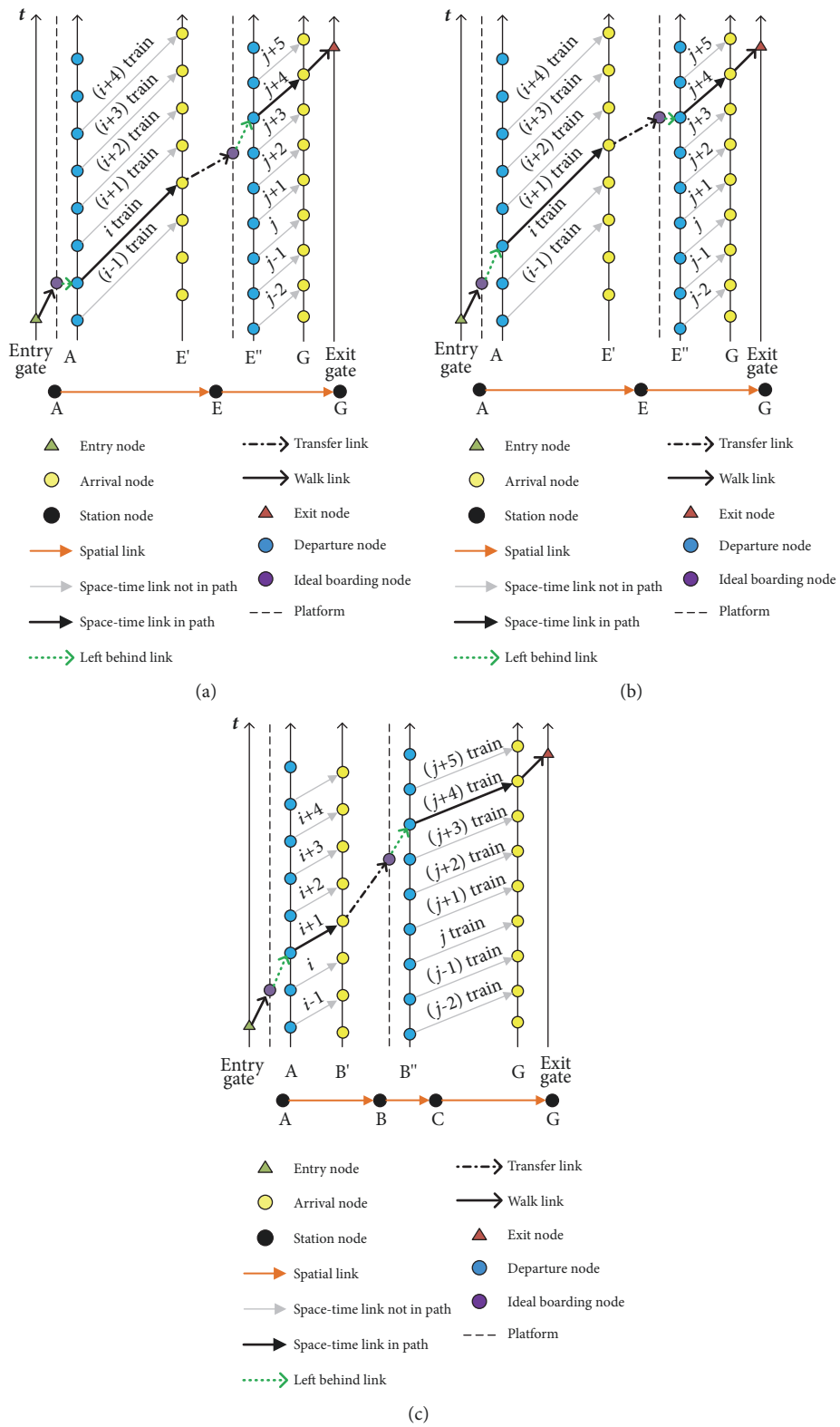

Figure 2: Feasible space-time-sequence trajectories.

Figure 2 presents three feasible space-time-sequence trajectories with given tap-in/tap-out constraints. This paper attempts to estimate the maximum-likelihood trajectory to mining passengers' travel patterns. Then, the problem of estimating the passenger's travel patterns can be converted into problems of generating feasible space-time-sequence trajectories and weight assignments.

4. Methodology

The key question in estimating a passenger's maximum space-time-sequence trajectory is how to generate a set of effective trajectories with entry and exit constraints and calculate their weights. Given AFC and AVL data, this section constructs a passenger's space-time-sequence trajectory-estimation model to maximize the weight of a chosen path. Tables 1 and 2 define the related notations and estimation variables used in the mathematical formulations.

4.1. Precise Estimation Model for a Passenger's Space-Time-Sequence Trajectory. Within a closed URT system, an itinerary begins with passing an entry gate and ends with swiping a smart card at an exit gate. The proposed model attempts to mine more detailed travel information about a passenger's travel using the tap-in and tap-out record. This paper addresses only regular passengers whose travels consist of some or all of the five activities present in Section 3. Some special circumstances such as a passenger forgetting to alight are not considered.

Considering that passengers are independent individuals, it is reasonable to assume that passengers' maximum-likelihood space-time-sequence trajectories are mutually independent. Furthermore, each passenger's walking activity and the trains running on different subway lines are independent from each other. Thus, the weight of each space-time-sequence trajectory is the product of probabilities of all space-time-sequence arcs passed by a passenger and the objective function is represented as follows.

$$\max z = \prod_{p \in P} \prod_{(i,j,t,t',k,k') \in E} \left(p_{i,t,k}^{j,t',k'} \right)^{y_{i,t,k,p}^{j,t',k'}} \tag{1}$$

subject to the following.

Space-Time-Sequence Flow Balance Constraints. If space-time-sequence node $(\boldsymbol{i}, \boldsymbol{t}, \boldsymbol{k})$ is an entry node, then

$$\sum_{(j,t',k') \in V} y_{i,t,k,p}^{j,t',k'} - \sum_{(j,t',k') \in V} y_{j,t',k',p}^{i,t,k} = 1 \tag{2}$$

If it is an exit node, then

$$\sum_{(j,t',k') \in V} y_{i,t,k,p}^{j,t',k'} - \sum_{(j,t',k') \in V} y_{j,t',k',p}^{i,t,k} = -1 \tag{3}$$

If $(\boldsymbol{i}, \boldsymbol{t}, \boldsymbol{k})$ is an intermediate node, then

$$\sum_{(j,t',k') \in V} y_{i,t,k,p}^{j,t',k'} - \sum_{(j,t',k') \in V} y_{j,t',k',p}^{i,t,k} = 0 \tag{4}$$

Table 1: Notations and input parameters.

Symbol	Definitions
$\mathbf{S}$	Set of URT stations
$\mathbf{N}$	Set of spatial nodes
$\mathbf{N^P}$	Set of platform spatial nodes, $\mathbf{N^P} \subset \mathbf{N}$
$\mathbf{C}$	Set of spatial connections, including sections and transfer links
$\mathbf{E}$	Set of space-time-sequence arcs
$\mathbf{E^W}$	Set of walking space-time-sequence arcs
$\mathbf{E^L}$	Set of left-behind space-time-sequence arcs
$\mathbf{V}$	Set of space-time-sequence nodes
$\mathbf{T}$	Set of activity times
$\mathbf{P}$	Set of passengers
$\boldsymbol{s}, \boldsymbol{s'}$	Index of urban railway transit stations, $\boldsymbol{s}, \boldsymbol{s'} \in \mathbf{S}$
$\mathbf{L}_i$	Set of trains passing platform $\boldsymbol{i}$, $\boldsymbol{i} \in \mathbf{N^P}$
$\mathbf{L}_i(\boldsymbol{k})$	The $\boldsymbol{k}$th train passing platform $\boldsymbol{i}$, $\boldsymbol{i} \in \mathbf{N^P}$
$\boldsymbol{p}$	Index of passenger, $\boldsymbol{p} \in \mathbf{P}$
$\boldsymbol{i}, \boldsymbol{j}$	Index of spatial nodes, $\boldsymbol{i}, \boldsymbol{j} \in \mathbf{N}$
$\boldsymbol{t}, \boldsymbol{t'}$	Index of time stamp, $\boldsymbol{t}, \boldsymbol{t'} \in \mathbf{T}$
$\boldsymbol{h}_s^t$	Interval time at station $\boldsymbol{s}$ at time $\boldsymbol{t}$, $\boldsymbol{s} \in \mathbf{S}, \boldsymbol{t} \in \mathbf{T}$
$\boldsymbol{s}_i$	The station to which node $\boldsymbol{i}$ belongs, $\boldsymbol{i} \in \mathbf{N}, \mathbf{s} \in \mathbf{S}$
$(\boldsymbol{i}, \boldsymbol{j})$	Index of spatial connection, $(\boldsymbol{i}, \boldsymbol{j}) \in \mathbf{C}$
$(\boldsymbol{i}, \boldsymbol{t}, \boldsymbol{k}), (\boldsymbol{j}, \boldsymbol{t'}, \boldsymbol{k'})$	Index of space-time-sequence node, $(\boldsymbol{i}, \boldsymbol{t}, \boldsymbol{k}), (\boldsymbol{j}, \boldsymbol{t'}, \boldsymbol{k}) \in \mathbf{V}$
$(\boldsymbol{i}, \boldsymbol{j}, \boldsymbol{t}, \boldsymbol{t'}, \boldsymbol{k}, \boldsymbol{k'})$	Index of space-time-sequence arc indicating departing $\boldsymbol{i}$ at $\boldsymbol{t}$ during $\boldsymbol{k}$ time interval and arriving at $\boldsymbol{j}$ at $\boldsymbol{t'}$ during $\boldsymbol{k'}$ time interval, $(\boldsymbol{i}, \boldsymbol{j}, \boldsymbol{t}, \boldsymbol{t'}, \boldsymbol{k}, \boldsymbol{k'}) \in \mathbf{E}$
$\boldsymbol{f}_{i,t,k}^{j,t',k\prime}$	The distribution of the time cost of space-time-sequence arc $(\boldsymbol{i}, \boldsymbol{j}, \boldsymbol{t}, \boldsymbol{t'}, \boldsymbol{k}, \boldsymbol{k'})$
$\boldsymbol{\mu}_{i,t,k}^{j,t',k\prime}$	The mean of the time cost of space-time-sequence arc $(\boldsymbol{i}, \boldsymbol{j}, \boldsymbol{t}, \boldsymbol{t'}, \boldsymbol{k}, \boldsymbol{k'})$
$\boldsymbol{\sigma}_{i,t,k}^{j,t',k\prime}$	The variance of the time cost of space-time-sequence arc $(\boldsymbol{i}, \boldsymbol{j}, \boldsymbol{t}, \boldsymbol{t'}, \boldsymbol{k}, \boldsymbol{k'})$
$\boldsymbol{p}_{i,t,k}^{j,t',k\prime}$	The probability that a passenger leaves node $\boldsymbol{i}$ at $\boldsymbol{t}$ and arrives at node $\boldsymbol{j}$ at $\boldsymbol{t'}$
$\boldsymbol{c}_{i,t,k}^{j,t',k\prime}$	The time that a passenger should spend on $(\boldsymbol{i}, \boldsymbol{j}, \boldsymbol{t}, \boldsymbol{t'}, \boldsymbol{k}, \boldsymbol{k'})$
$\boldsymbol{l}_{i,t,k}^{j,t',k\prime}$	The number of times passengers who travel by $(\boldsymbol{i}, \boldsymbol{j}, \boldsymbol{t}, \boldsymbol{t'}, \boldsymbol{k}, \boldsymbol{k'})$ are left-behind
$\boldsymbol{t}_{i,k}^{a}, \boldsymbol{t}_{i,k}^{d}$	The arrival and departure times at platform $\boldsymbol{i}$ of the $\boldsymbol{k}$**th** train, $\boldsymbol{i} \in \mathbf{N^P}$
$\boldsymbol{\varepsilon}$	The upper error limit of function value
$\boldsymbol{I}_{i,k}$	The $\boldsymbol{k}$ interval time of platform $\boldsymbol{i}$, $\boldsymbol{i} \in \mathbf{N^P}$

Constraints (2)-(4) ensure flow balance at the network entry, exit and intermediate space-time-sequence nodes, respectively.

On-Train Constraints. If $(\boldsymbol{i}, \boldsymbol{j}, \boldsymbol{t}, \boldsymbol{t'}, \boldsymbol{k}, \boldsymbol{k'})$ indicates on-train activity, the train at the start point should be the same as that

Table 2: Estimation variables.

Variable	Definition
μ_s^a	The mean of access walking time at station s, $s \in \mathbf{S}$.
σ_s^a	The variance of access walking time at station s, $s \in \mathbf{S}$.
μ_s^e	The mean of egress walking time at station s, $s \in \mathbf{S}$.
σ_s^e	The variance of egress walking time at station s, $s \in \mathbf{S}$.
$\mu_{s,s\prime}^c$	The mean time consumed transferring from s to s', s, $s' \in \mathbf{S}$.
$\sigma_{s,s\prime}^c$	The variance of time consumed transferring from s to s', s, $s' \in \mathbf{S}$.
μ_s^w	The mean of platform waiting time at station s during off-peak hours, $s \in \mathbf{S}$.
σ_s^w	The variance of platform waiting time at station s during off-peak hours, $s \in \mathbf{S}$.
$\mu_{i,j}$	The mean number of times passengers are left-behind on platform i.
$y_{i,t,k,p}^{j,t',k\prime}$	0-1 binary variables: 1 if trajectory of passenger p contains space-time-sequence arc (i, j, t, t', k, k'); 0 otherwise.

of the end point. If the train runs in the upward direction, then

$$\mathbf{L}_i(k) = \mathbf{L}_j\left(k'\right) \tag{5}$$

Sequence Constraints. According to the definition of a space-time-sequence, if (i, t, k) is an ideal boarding node or train departure node, then

$$t = t_{i_{station},k}^d \tag{6}$$

or else

$$t_{i_{station},k-1}^d \le t < t_{i_{station},k}^d \tag{7}$$

Note that the sequence number of the end node of a space-time-sequence arc should not be less than that of the start node. Thus, if (i, j, t, t', k, k') is a left-behind space-time-sequence arc, then

$$k \le k' \tag{8}$$

4.2. Calculation of Weight of Space-Time-Sequence Arc. According to the objective function, the precise estimation problem for a passenger's detailed travel patterns is converted into a problem of the generation and weight assignments of a feasible trajectory set. Among a passenger's travel activities within the URT system, access, egress, and transfer are walking activities. Once passengers board a train, their movements with respect to time are the same as those of train vehicles and correspond to AVL data. Thus, if (i, j, t, t', k, k') is an on-train space-time-sequence arc, $p_{i,t,k}^{j,t',k\prime}$ should be equal to 1.

$$p_{i,t,k}^{j,t',k\prime} = 1 \tag{9}$$

According to Shi [27], the distribution of passengers' walking times is similar to a normal distribution. In other words, the distributions of access walking time, egress walking time, and transfer walking time are also similar to normal distributions. An access space-time-sequence arc(i, j, t, t', k, k') means that the passenger departs node i at time t and arrives node j before time t'. Thus, $p_{i,t,k}^{j,t',k\prime}$ can be calculated as follows.

$$p_{i,t,k}^{j,t',k'} = \begin{cases} \int_0^{t'-t} \frac{1}{\sqrt{2\pi}\sigma_{i,t,k}^{j,t',k'}} \exp\left(-\frac{\left(x-\mu_{i,t,k}^{j,t',k'}\right)^2}{2\left(\sigma_{i,t,k}^{j,t',k'}\right)^2}\right) dx, & k = k' \\ \int_{t'-t-I_{j,k'}}^{t'-t} \frac{1}{\sqrt{2\pi}\sigma_{i,t,k}^{j,t',k'}} \exp\left(-\frac{\left(x-\mu_{i,t,k}^{j,t',k'}\right)^2}{2\left(\sigma_{i,t,k}^{j,t',k'}\right)^2}\right) dx, & k \ne k' \end{cases} \tag{10}$$

Similarly, if (i, j, t, t', k, k') is a transfer space-time-sequence arc, its probability can also be calculated by (10). Equation (10) can be integrated using characteristics of the normal distribution, and the result is given as follows.

$$p_{i,t,k}^{j,t',k'} = \begin{cases} \Phi\left(\frac{t'-t-\mu_{i,t,k}^{j,t',k'}}{\sigma_{i,t,k}^{j,t',k'}}\right) - \Phi\left(\frac{-\mu_{i,t,k}^{j,t',k'}}{\sigma_{i,t,k}^{j,t',k'}}\right), & k = k' \\ \Phi\left(\frac{t'-t-\mu_{i,t,k}^{j,t',k'}}{\sigma_{i,t,k}^{j,t',k'}}\right) - \Phi\left(\frac{t'-t-I_{j,k'}-\mu_{i,t,k}^{j,t',k'}}{\sigma_{i,t,k}^{j,t',k'}}\right), & k \ne k' \end{cases} \tag{11}$$

In (11), $\Phi(x)$ is the standard normal distribution function; its numerical values can be found in the tables.

If (i, j, t, t', k, k') is a left-behind space-time-sequence arc, the number of times the passenger is left-behind and can be calculated as follows.

$$l_{i,t,k}^{j,t',k\prime} = k' - k \tag{12}$$

Because that passenger can board the first train after their arrival at a platform during off-peak hours, the number of times passengers is left-behind and $\mu_{i,j}$ should be zero. Similar to the on-train space-time-sequence arc, $p_{i,t,k}^{j,t',k\prime}$ should be equal to 1 during off-peak hours.

$$p_{i,t,k}^{j,t',k\prime} = 1 \tag{13}$$

and

$$k' = k \tag{14}$$

Provided that all passengers in line board the train following the first-come-first-served principle during peak hours, the closer the number of left-behinds is to its mean value, the greater the probability $p_{i,t,k}^{j,t',k\prime}$ is. It is similar to the progressive distribution. Here, we adopt (15) as its distribution function.

$$p_{i,t,k}^{j,t',k\prime} = \frac{1}{\sqrt{2\pi}} e^{-\left(l_{i,t,k}^{j,t',k\prime}-\mu_{i,j}\right)^2/2} \tag{15}$$

4.3. Estimation of Station Walking-Time Parameters

Estimation of Platform Waiting-Time Parameters. Platform waiting time is defined as the elapsed time after a passenger's arrival at a platform and before the departure time of the first departing train. During off-peak hours, the capacity of the transit service supply is greater than traffic demand, and all passengers can board the first departing train after their arrival at a platform. The actual boarding nodes are the same as the ideal boarding nodes in this situation. Additionally, the platform waiting time should be less than the interval between the departure times of sequential trains passing the station. Since the times passengers arrive at a platform are random, the distribution at the platform is the uniform distribution. The mean and variance of the platform waiting-time distribution are given as follows.

$$\mu_s^w = \frac{h_s^t}{2} \tag{16}$$

$$\left(\sigma_s^w\right)^2 = \frac{\left(h_s^t\right)^2}{12} \tag{17}$$

Estimation of Station Walking-Time Parameters. According to Section 2, there are three types of walking activities at a station: access at an origin station, egress at the destination station, and transfer(s) at transfer station(s). If (i, j, t, t', k, k') is an exit space-time-sequence arc, $c_{i,t,k}^{j,t',k'}$ is equal to the egress walking time. Thus, it has the same distribution function. The mean and variance of the time consumption of (i, j, t, t', k, k') can be calculated as follows.

$$\mu_{i,t,k}^{j,t',k'} = \mu_s^e \tag{18}$$

$$\left(\sigma_{i,t,k}^{j,t',k'}\right)^2 = \left(\sigma_s^e\right)^2 \tag{19}$$

where $s_i = s_j = s$, and (j, t', k') is an exit space-time-sequence node.

If (i, j, t, t', k, k') is an entry space-time-sequence arc and its time consumption consists of two parts, (i) access walking time and (ii) platform waiting time, thus, the mean and variance of the access walking-time distribution can be calculated by

$$\mu_s^a = \mu_{i,t,k}^{j,t',k'} - \mu_s^w \tag{20}$$

$$\left(\sigma_s^a\right)^2 = \left(\sigma_{i,t,k}^{j,t',k'}\right)^2 - \left(\sigma_s^w\right)^2 \tag{21}$$

where $s_i = s_j = s$ and (i, t, k) is an entry space-time-sequence node.

Similarly, the parameters of the distribution of transfer walking times can be calculated as follows.

$$\mu_{s,s'}^c = \mu_{i,t,k}^{j,t',k'} \tag{22}$$

$$\left(\sigma_{s,s'}^c\right)^2 = \left(\sigma_{i,t,k}^{j,t',k'}\right)^2 \tag{23}$$

where $s_i = s$, $s_j = s'$ and $s(name) = s'(name)$.

5. Solution Algorithm

The objective function presented in Section 4 is a product and is nonlinear; it is difficult to optimize. This section presents an algorithm to estimate the maximum space-time-sequence trajectory for all passengers with given AFC data.

The objective function can be converted to a sum using properties of the logarithm function.

$$\max \ln z = \sum_{p \in P} \sum_{(i,j,t,t',k,k') \in E^W \cup E^L} y_{i,t,k,p}^{j,t',k'} \bullet \ln p_{i,t,k}^{j,t',k'} \tag{24}$$

By substituting (10)-(12), (14), and (16) into (24) we get

$$\begin{aligned} &\max \ln z \\ &= \sum_{p \in P} \sum_{(i,j,t,t',k,k') \in E^W} y_{i,t,k,p}^{j,t',k'} \bullet \ln\left(\Phi\left(\frac{t' - t - \mu_{i,t,k}^{j,t',k'}}{\sigma_{i,t,k}^{j,t',k'}}\right) - \Phi\left(\frac{\max\left\{0, t' - t - I_{j,k'}\right\} - \mu_{i,t,k}^{j,t',k'}}{\sigma_{i,t,k}^{j,t',k'}}\right)\right) - \sum_{p \in P} \sum_{(i,j,t,t',k,k') \in E^L} y_{i,t,k,p}^{j,t',k'} \bullet \left(l_{i,t,k}^{j,t',k'} - \mu_{i,j}\right)^2 - \sum_{p \in P} \sum_{(i,j,t,t',k,k') \in E^L} \ln \sqrt{2\pi} \end{aligned} \tag{25}$$

Provided that the mean and variance of a walking space-time-sequence arc, $\Phi((t' - \mu_{i,t,k}^{j,t',k'})/\sigma_{i,t,k}^{j,t',k'})$, and $\Phi((\max\{0, t' - t - I_{j,k'}\} - \mu_{i,t,k}^{j,t',k'})/\sigma_{i,t,k}^{j,t',k'})$ are fixed and can be calculated easily, similarly, $(l_{i,t,k}^{j,t',k'} - \mu_{i,j})^2$ can also be calculated. Thus, the weights can be calculated for all trajectories. Inverting this argument, if all passengers' space-time-sequence trajectories are given, these trajectories can be used as samples to estimate these parameters with the maximum-likelihood estimation algorithm. To solve the station time parameters estimation problem and a passenger's trajectory-estimation problem, an iterative optimization algorithm is proposed in this paper.

The algorithm assigns a feasible trajectory for all passengers randomly and estimates station time parameters with MLE. The estimation of passengers' maximum-likelihood space-time-sequence trajectories and station time parameters will not stop until the optimal result is obtained.

Algorithm 1 (iterative optimization algorithm). *Input:* passengers' feasible space-time-sequence trajectory sets

Output: station time parameters and passengers' maximum-likelihood space-time-sequence trajectories

Step 1 (initialization). Input AFC data and AVL data, initialize parameters of algorithm. Set $(\ln z)^0 = 0$, $\varepsilon = 10$, $k = 1$, where k is the index of calculation generation.

Step 2 (feasible space-time-sequence trajectories generation). Generate a feasible space-time-sequence trajectory set and select a trajectory randomly and assign it to all passengers.

Step 3 (estimate station parameters using the maximum-likelihood estimation algorithm). Given the detailed information of all passengers' space-time-sequence trajectories, a passenger's time consumption at any station can be calculated. Then, the mean and variance can be estimated according to (27) and (28).

$$\mu_{i,t,k}^{j,t',k'} = \frac{\sum_{p\in \mathrm{P}} \sum_{(i,j,t,t',k,k')\in \mathrm{E^W}} y_{i,t,k,p}^{j,t',k'} \bullet c_{i,t,k}^{j,t',k'}}{\sum_{p\in \mathrm{P}} \sum_{(i,j,t,t',k,k')\in \mathrm{E^W}} y_{i,t,k,p}^{j,t',k'}} \tag{26}$$

$$\left(\sigma_{i,t,k}^{j,t',k'}\right)^2 = \frac{\sum_{p\in \mathrm{P}} \sum_{(i,j,t,t',k,k')\in \mathrm{E^W}} y_{i,t,k,p}^{j,t',k'} \bullet \left(c_{i,t,k}^{j,t',k'} - \mu_{i,t,k}^{j,t',k'}\right)}{\sum_{p\in \mathrm{P}} \sum_{(i,j,t,t',k,k')\in \mathrm{E^W}} y_{i,t,k,p}^{j,t',k'}} \tag{27}$$

During peak hours, the mean of left-behind can also calculated.

$$\mu_{i,j} = \frac{\sum_{p\in \mathrm{P}} \sum_{(i,j,t,t',k,k')\in \mathrm{E^L}} y_{i,t,k,p}^{j,t',k'} \bullet l_{i,t,k}^{j,t',k'}}{\sum_{p\in \mathrm{P}} \sum_{(i,j,t,t',k,k')\in \mathrm{E^L}} y_{i,t,k,p}^{j,t',k'}} \tag{28}$$

Step 4 (deviation calculation). Calculate $(\ln \boldsymbol{z})^k$ according to (25). If $(\ln \boldsymbol{z})^{\boldsymbol{k}} - (\ln \boldsymbol{z})^{\boldsymbol{k}-1} < \varepsilon$, then go to Step 6. Otherwise, set $\boldsymbol{k} = \boldsymbol{k} + 1$ and go to Step 5.

Step 5 (passenger's travel pattern estimation). Assign the most likely space-time-sequence trajectory for all passengers by executing the passenger's travel-patterns estimation algorithm. Then, go to Step 3.

Step 6 (algorithm end). Calculate station time parameters based on (16)-(23). Output the station time parameters and all passengers' space-time-sequence trajectories.

We implemented the following iterative procedure to solve the problems, as summarized in Figure 3.

As shown in Figure 3, we have extended the Beijing rail transit network topology by replacing a station with four types of spatial nodes (entry, exit, platform, and track) from which we generate the set of feasible space-time-sequence trajectories using the methodology in [7]. The set of passengers' feasible space-time-sequence trajectories is the basis for solving the station time parameters estimation problem and the passenger's space-time-sequence trajectory-estimation problem. The mean and variance of passengers' time consumption at a station can be estimated with MLE once all passengers' space-time-sequence trajectories are assigned. Then, the key to the station time parameters estimation problem is the passenger assignment problem. After estimating station time parameters, all space-time-sequence arc time-consumption distribution functions are known. The weights of all space-time-sequence arcs are fixed and can be calculated. Thus, the passenger's travel-patterns estimation problem can be converted into a shortest-path problem. The passenger's travel-patterns estimation algorithm is presented as follows.

Algorithm 2 (passenger's travel-patterns estimation algorithm). *Input*: station time parameters

Output: passengers' space-time-sequence trajectories

Step 1 (initialize algorithm parameters). Set $\boldsymbol{k} = 1$, $\boldsymbol{n}$ is the count of passengers.

Step 2 (weight calculation). Update weights of all space-time-sequence arcs according to the following equations.

$$\ln p_{i,t,k}^{j,t',k'} = \begin{cases} -\dfrac{\left(c_{i,t,k}^{j,t',k'} - \mu_{i,t,k}^{j,t',k'}\right)^2}{2\left(\sigma_{i,t,k}^{j,t',k'}\right)^2} + \ln \dfrac{1}{\sqrt{2\pi}\sigma_{i,t,k}^{j,t',k'}}, & \left(i,j,t,t',k,k'\right) \in \mathrm{E^W} \\ -\left(l_{i,t,k}^{j,t',k'} - \mu_{i,j}\right)^2, & \left(i,j,t,t',k,k'\right) \in \mathrm{E^L} \\ 0, & otherwise \end{cases} \tag{29}$$

Step 3 (estimate a passenger's maximum-likelihood space-time-sequence trajectory). Calculate the weight of each feasible space-time-sequence trajectory following (25) and assign the shortest path for the $\boldsymbol{k}$th passenger.

Step 4 (algorithm loop judgement). If $\boldsymbol{k} > \boldsymbol{n}$, go to Step 4; otherwise, go to Step 2.

Step 5 (algorithm end). Output all passengers' space-time-sequence trajectories.

6. Case Study

To verify the proposed methodology, the model was tested on real-world data from the Beijing railway transit network. We propose an approach to validation that addresses the lack of actual passengers' itineraries. The approach is to analyze the estimation results statistically and compare them with existing results. A software system has been developed using C#, Windows Presentation Foundation (WPF) and Human-computer interaction technology based on the model proposed in this paper.

6.1. Beijing Railway Transit Network. Construction started on the Beijing railway transit network in 1965, and the first line began operation on January 15, 1971. By the end of 2016, Beijing railway transit had developed into a large-scale network, and its average daily passenger traffic was close to 10 million. During this period, the metro operations were divided into two companies: (i) Beijing Subway and (ii) Beijing MTR. Figure 4 shows the real-world Beijing railway transit network topology at the end of 2016.

As shown in Figure 4, there are a total of 17 railway lines and 338 stations, including 53 transfer stations. Unlike other railway lines, the airport line, labeled JC in Figure 4, is

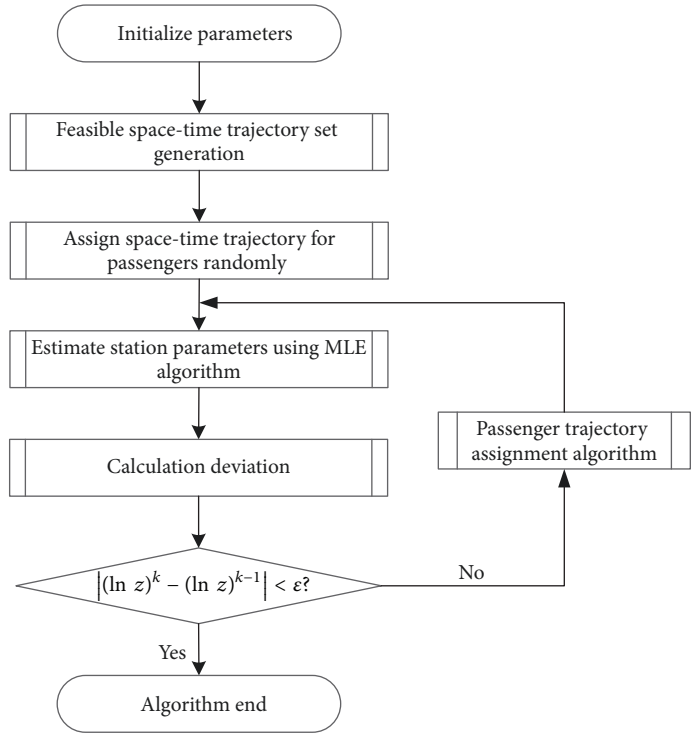

Figure 3: Solution methodology procedure.

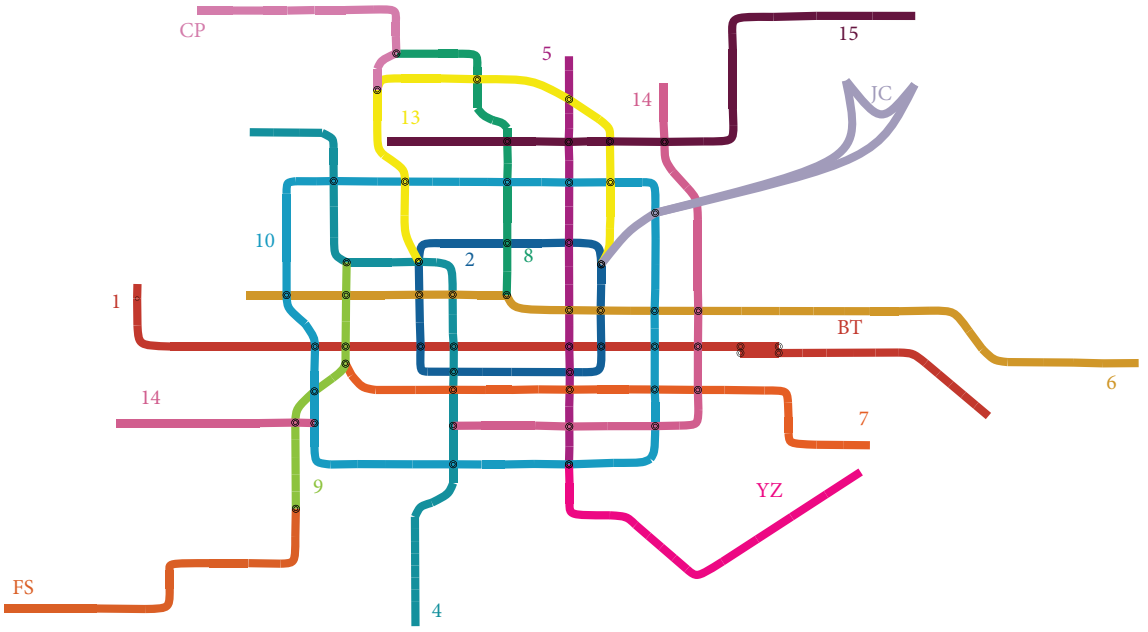

Figure 4: Topology of Beijing rail transit network.

independent; passengers must swipe their smart cards when they transfer from/to the airport line.

6.2. Input Data Preparation. The case study employs AFC transaction data and AVL data obtained from Beijing Subway and Beijing MTR Corporation. Although these data vary from manufacturer to manufacturer, the basic information used in this paper was recorded.

An automatic ticketing system was launched in Beijing railway transit on June 9, 2008. Dozens of types of information are recorded and saved to a database. Passengers' tap-in and tap-out information was extracted, and Table 3

Table 3: AFC transaction data from Dongwuyuan to Chongwenmen.

CardID	Entry Time	Exit Time
15093414604	17:34:00	18:08:16
15094844706	18:09:00	18:43:23
15094957624	18:08:00	18:42:11
15095055101	17:29:00	18:01:21
15094845984	17:47:00	18:17:21
15095014893	17:27:00	17:56:18
15093507376	17:25:00	17:54:17
15093510397	17:43:00	18:12:17
15095109950	18:04:00	18:33:02

shows some AFC data from Dongwuyuan to Chongwenmen observed during peak hours on May 9, 2016. The time format is hh:mm:ss.

Five types of information are extracted directly from the database. The Card ID is the unique identifier of a smart card and represents an individual passenger. It is critical to match tap-in and tap-out records. The time and location of tap-in and tap-out are recorded by the AFC system when passengers pass the entry/exit gates and swipe their smart cards. The data are accurate and record to the nearest second.

AVL data is actual train operation information; arrival and departure times at each station of all trains are recorded in detail, and then they are collected by tracking systems or CBTC, generally in one of two formats within the Beijing railway transit (see Tables 4(a) and 4(b)).

The software system uses a unified data structure for trains to analyze passengers' travel behavior at the network level and improve computation efficiency.

6.3. Estimation Results and Discussion. The accuracy of the estimation results is the key measure in evaluating the model proposed in this paper. Accuracy is evaluated at both the individual and network levels.

6.3.1. Estimation Results for Station Parameters. Station-time parameters are the basic attributes that tell the approximate time consumed, including access, egress, and transfer walking times for a transfer station. A maximum-likelihood estimation algorithm was developed to estimate these station time parameters by analyzing passengers' space-time-sequence trajectories. Figure 6 shows the distribution of station-time parameters during off-peak hours and peak hours. The distributions of access walking time in Dongwuyuan station during off-peak hours and peak hours are shown in Figures 5(a) and 5(b). Figures 5(c) and 5(d) show the distribution of egress walking time and the last two figures present the distribution of transfer walking time.

As shown in Figure 5, the distributions are obviously similar to a normal distribution. The actual times consumed are concentrated over a certain range. Table 5(a) compares the estimated access walking time with a manual survey result provided by the Beijing Transportation Operations Coordination Center (TOCC). Tables 5(b) and 5(c) show comparisons of egress and transfer walking times.

According to Tables 5(a), 5(b), and 5(c), the estimation results are close to the manual survey results. The relative deviations between estimation results and manual survey are less than 5% except for the access walking time of Dongwuyuan. As the access walking time is approximately 30 seconds, the absolute deviation is only three or four seconds and is acceptable, although the relative deviation exceeds 5%.

6.3.2. Left-Behind. The distribution of the estimated number of left-behinds is shown in Figure 6. During off-peak hours, only approximately two percent of passengers are left-behind, and the vast majority of passengers board the first train. These left-behind passengers may have been waiting for a companion. Most passengers left-behind during peak hours due to train vehicle capacity constraints miss only one train, though some passengers miss as many as three.

6.3.3. Passenger Travel Patterns. A space-time-sequence indicates a passenger's travel information in detail. The number of times the passenger is left-behind and the specific train(s) taken can be found directly. Figure 7 shows the space-time-sequence trajectory-estimation result of a passenger whose Card ID is 15093414604. This passenger passed the Dongwuyuan entry gate at 17:34:00 and left from Chongwenmen at 18:08:16. Table 6 shows the station-time parameters estimated in Sections 6.3.1 and 6.3.2.

As shown in Figure 7, there are theoretically more than ten feasible space-time-sequence trajectories. Partial trajectories are given in Table 7.

In Table 7, columns 3-5 indicate the total time consumed with no left-behinds at the origin, transfer, and destination stations. The total time consists of two components at both the origin and transfer stations: the access/transfer and platform waiting times. In general, the total time consumed at the destination station should be the egress time. According to the estimation, this passenger chose the fourth trajectory for the journey.

The passenger arrived at platform of Dongwuyuan before 17:35:28, the departure time of train 1S443. He/she did not leave Dongwuyuan by 1S443 until the departure of the next train, 1Q445, because of crowding. Similarly, he/she was left-behind once at Xizhimen and boarded train 322223 for his/her destination.

6.3.4. Distribution of URT Network Passenger Flow. The distribution of the URT network passenger flow is one of most important network characteristics for URT operations. It reflects the time-dependent travel demand and is the basis for the transportation plan. Figure 8 shows the distribution of URT network passenger flow during peak hours.

During peak hours, as shown in Figure 8, the URT network is crowded, and the maximum train load is up to 140%. The more congested sections are located mainly in the center of Beijing, consistent with that during off-peak hours [7]. Table 8 compares the estimated results of section passenger flow and the results provided by Beijing TOCC for the top five subway sections during peak hours.

As shown in Table 8, the estimated results are consistent with the results from TOCC. Compared with off-peak hours,

Table 4

(a) AVL data of Line 2

Train Num	Xizhimen		Xuanwumen		Chongwenmen	
	Arrive Time	Depart Time	Arrive Time	Depart Time	Arrive Time	Depart Time
302108	10:12:30	10:13:30	10:24:31	10:25:01	10:31:07	10:31:52
322109	10:16:00	10:17:00	10:28:01	10:28:31	10:34:37	10:35:22
192110	10:20:30	10:21:30	10:32:31	10:33:01	10:39:07	10:39:52
362111	10:25:00	10:26:00	10:37:01	10:37:31	10:43:37	10:44:22
382112	10:29:30	10:30:30	10:41:31	10:42:01	10:48:07	10:48:52
402113	10:34:00	10:35:00	10:46:01	10:46:31	10:52:37	10:53:22
62114	10:38:30	10:39:30	10:50:31	10:51:01	10:57:07	10:57:52

(b) AVL data of Line 10

Rail Line	Train Num	Station	Arrival Time	Departure Time
...	...	...	...	...
10	2278	Jinsong	17:48:50	17:49:25
10	2278	Shuangjing	17:50:52	17:51:22
10	2278	Guomao	17:53:30	17:54:20
...	...	...	...	...
10	2302	Liuliqiao	19:44:20	19:45:15
10	2302	Xiju	19:47:06	19:47:51
10	2302	Niwa	19:49:01	19:49:31

Table 5

(a) Dongwuyuan access walking-time parameter comparison table

	Survey (s)	Estimation (s)	Relative deviation
Peak	34	37	8.82%

(b) Chongwenmen egress walking-time parameter comparison table

	Survey (s)	Estimation (s)	Relative deviation
Peak	169	177	4.73%

(c) Transfer walking-time parameter comparison table

Station	Transfer Direction	Survey (s)	Estimation (s)	Relative deviation
Xizhimen	Line 4⟶ Line 2	219	225	2.74%
Xuanwumen	Line 4 ⟶ Line 2	240	229	4.58%

Table 6: Station parameters expectation at each subway station.

Station parameter	Expected value
Access time at Dongwuyuan	37 s
Transfer time from Line 4 to Line 2 at Xizhimen	225 s
Egress time from Chongwenmen	169 s
Left-behind at Dongwuyuan	1
Left-behind at Xizhimen	1

the distribution of relatively congested sections during peak hours is similar. The relatively congested sections mainly focus on the out-of-Beijing direction and are located around Central Business Districts (CBD) during off-peak and peak. The difference is that a large number of passengers get off work and travel to the suburbs through the URT network during late peak hours.

7. Conclusion

As the objective of URT service, passenger flow is the basis for the URT transportation organization. The characteristics of the passenger flow distribution have a large impact on the transportation plan and efficiency. To better understand the composition of passenger flow and its characteristics, we have developed a data-driven approach to estimate passengers' travel pattern. A space-time-sequence trajectory model was constructed based on AFC transaction data and AVL data to simulate passengers' travel processes.

An iterative maximum-likelihood estimation algorithm is presented to estimate the most likely space-time-sequence trajectory and station time parameters. A space-time-sequence trajectory indicates a passenger's travel activities and time consumed. Additionally, the number of left-behinds can be calculated by analyzing the relationship between a passenger's arrival time at a platform and the train sequence. The space-time-sequence trajectory is useful in mining passengers' path-choice behaviors and further assists URT operations to better predict the future operations status of the URT network.

Our future research will focus on two major areas: first, mining passengers' travel pattern in cases of emergency and analyzing the impact of an emergency on passengers' path choices and the URT network passenger flow distribution; second, how to optimize the URT timetable based on passenger travel demand and rescheduling in emergencies.

Table 7: Detailed information of theoretical feasible space-time-sequence trajectories.

No.	Board Train	Origin	Transfer	Destination	Left-behind
1	1S443, 422219	88 s	65 s	676 s	none
2	1Q445, 162221	88 s	205 s	489 s	once at Dongwuyuan
3	1Q445, 022222	88 s	205 s	339 s	once each at Dongwuyuan and Xizhimen
4	1Q445, 322223	88 s	355 s	189 s	once each at Dongwuyuan and Xizhimen
5	1A447, 322223	88 s	235 s	189 s	twice at Dongwuyuan and once at Xizhimen

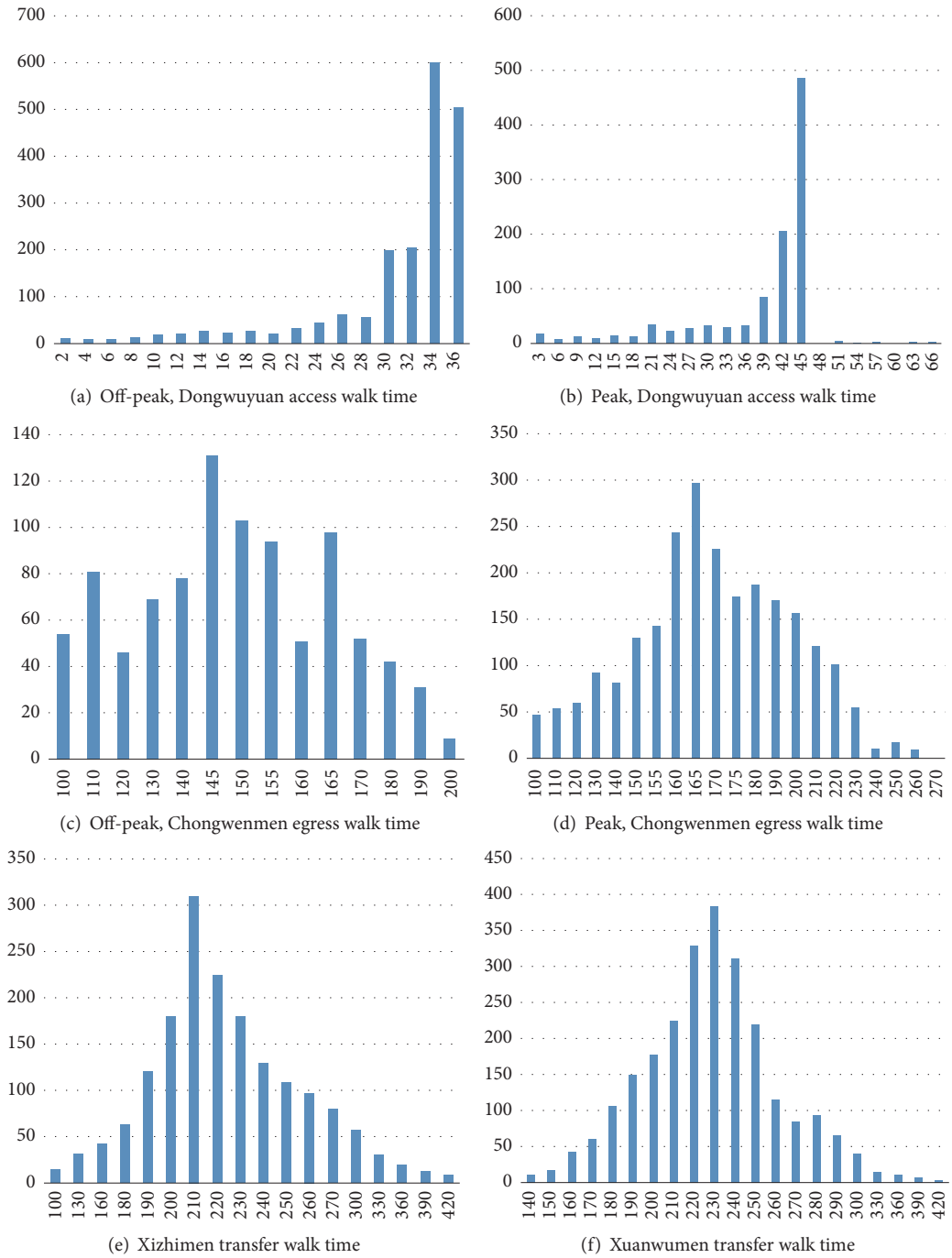

Figure 5: Estimation result of distribution of station-time parameters.

Table 8: Section passenger-flow comparison results.

Section name	TOCC	Estimate	Relative deviation
Xuanwumen - Caishikou	33458	33392	-0.20%
Caishikou – Taoranting	31898	31890	-0.03%
Jintailu – shilipu	31502	31576	0.23%
Hujialou – Jintailu	30171	30287	0.38%
Taoranting – Beijing South	28638	28651	0.03%

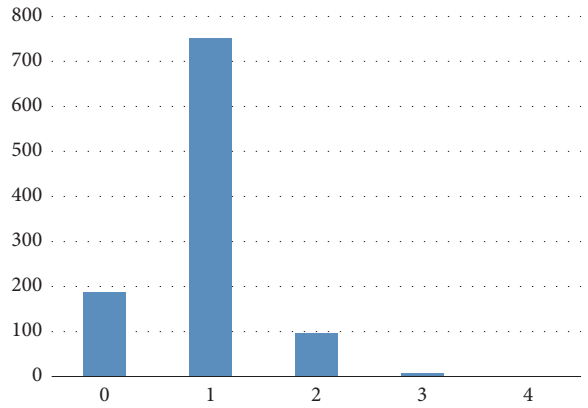

Figure 6: Distribution of number of left-behinds at Dongwuyuan.

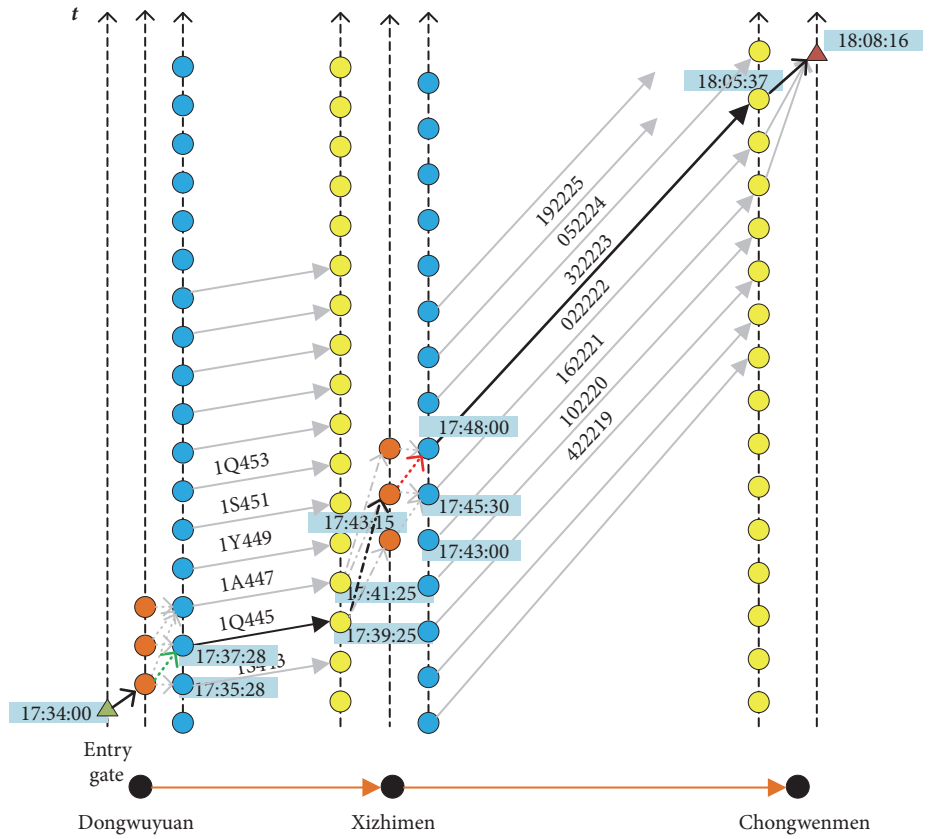

Figure 7: Space-time-sequence trajectory-estimation result.

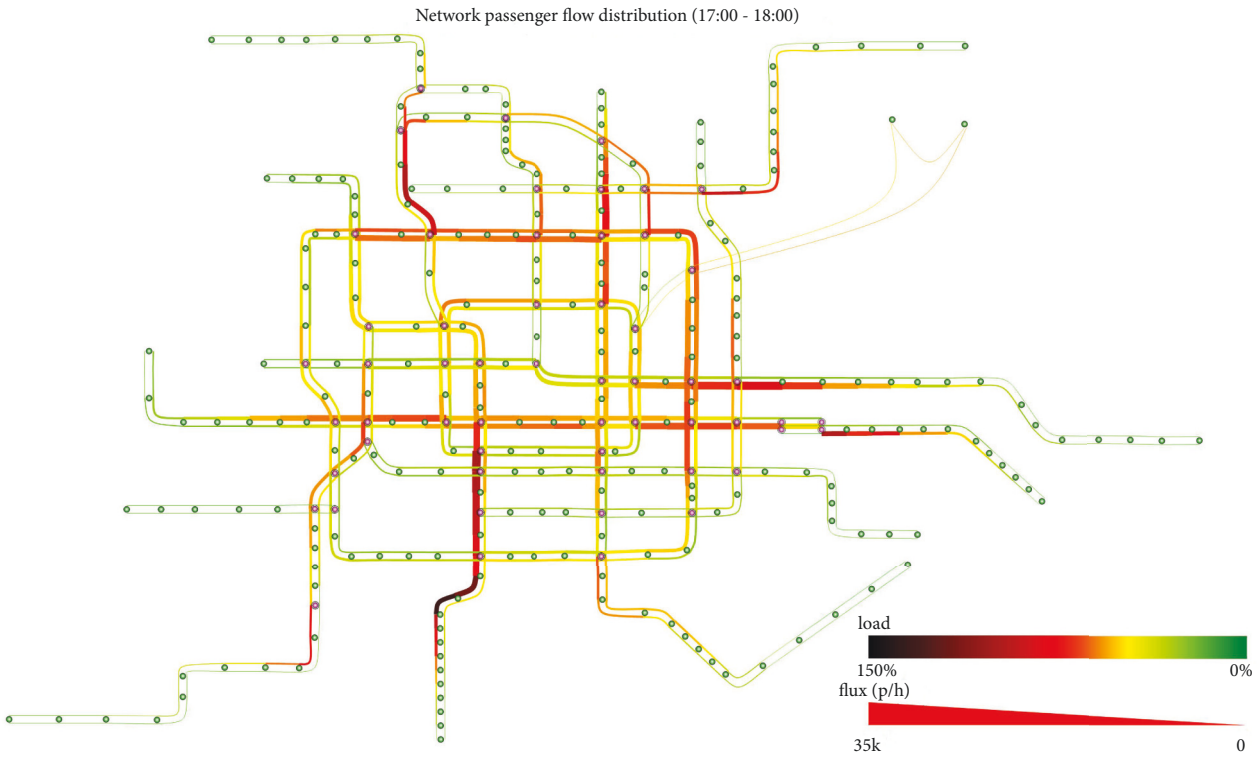

Figure 8: Distribution of URT network passenger flow during peak hours.

Data Availability

The data used to support the findings of this study have not been made available because AFC transaction data and AVL data are recorded during actual operations. These data can be mined for too much information and relate to passenger travel privacy as well as traffic safety. Readers can access the distribution of Beijing URT network passenger flow at https://map.bjsubway.com/. If necessary, we can process the data and provide it.

Disclosure

The data in this paper are based on research supported by the Beijing TOCC.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was financially supported by the Fundamental Research Funds (Grant T17JB00040), the Fundamental Research Funds for the Central Universities (Grant 2018RC012), and National Key R&D Program of China (2018YFB1201402).

References

[1] "Urban rail transit statistical analysis report," http://www.camet.org.cn.

[2] "National bureau of statistics of China," http://www.stats.gov.cn/.

[3] "Transport department of the government of the hong kong special administrative region," http://www.td.gov.hk.

[4] J. F. Guan, H. Yang, and S. C. Wirasinghe, "Simultaneous optimization of transit line configuration and passenger line assignment," *Transportation Research Part B: Methodological*, vol. 40, no. 10, pp. 885–902, 2006.

[5] B. Yu, Z. Yang, C. Cheng, and C. Liu, "Optimizing bus transit network with parallel ant colony algorithm," in *Proceedings of the Eastern Asia Society for Transportation Studies*, vol. 5, pp. 374–389, 2005.

[6] H. Niu and X. Zhou, "Optimizing urban rail timetable under time-dependent demand and oversaturated conditions," *Transportation Research Part C: Emerging Technologies*, vol. 36, pp. 212–230, 2013.

[7] X. Chen, L. Zhou, Y. Yue et al., "Data-driven method to estimate the maximum likelihood space-time trajectory in an Urban Rail Transit system," *Sustainability*, vol. 10, no. 6, article 1752, 2018.

[8] C. O. Tong and S. C. Wong, "A stochastic transit assignment model using a dynamic schedule-based network," *Transportation Research Part B: Methodological*, vol. 33, no. 2, pp. 107–121, 1999.

[9] C. O. Tong and S. C. Wong, "A schedule-based time-dependent trip assignment model for transit networks," *Journal of Advanced Transportation*, vol. 33, no. 3, pp. 371–388, 1999.

[10] J. Moller-Pedersen, "Assignment model for timetable based systems (TPSCHEDULE)," in *Proceedings of 27th European Transportation Forum, Seminar F*, pp. 159–168, Cambridge, England, 1999.

[11] O. A. Nielsen and G. Jovicic, "A large-scale stochastic timetable-based transit assignment model for route and submode choices," in *Proceedings of 27th European Transportation Forum, Seminar F*, pp. 169–184, Cambridge, England, 1999.

[12] R.-H. Xu, Q. Luo, and P. Gao, "Passenger flow distribution model and algorithm for urban rail transit network based on multi-route choice," *Journal of the China Railway Society*, vol. 31, no. 2, pp. 110–114, 2009.

[13] Y. Hamdouch and S. Lawphongpanich, "Schedule-based transit assignment model with travel strategies and capacity constraints," *Transportation Research Part B: Methodological*, vol. 42, no. 7-8, pp. 663–684, 2008.

[14] M. E. Ben-Akiva, S. Gao, Z. Wei, and Y. Wen, "A dynamic traffic assignment model for highly congested urban networks," *Transportation Research Part C: Emerging Technologies*, vol. 24, pp. 62–82, 2012.

[15] M. Cepeda, R. Cominetti, and M. Florian, "A frequency-based assignment model for congested transit networks with strict capacity constraints: characterization and computation of equilibria," *Transportation Research Part B: Methodological*, vol. 40, no. 6, pp. 437–459, 2006.

[16] R. D. Connors and A. Sumalee, "A network equilibrium model with travellers' perception of stochastic travel times," *Transportation Research Part B: Methodological*, vol. 43, no. 6, pp. 614–624, 2009.

[17] J.-D. Schmöcker, M. G. H. Bell, and F. Kurauchi, "A quasi-dynamic capacity constrained frequency-based transit assignment model," *Transportation Research Part B: Methodological*, vol. 42, no. 10, pp. 925–945, 2008.

[18] C. O. Tong and S. C. Wong, "A predictive dynamic traffic assignment model in congested capacity-constrained road networks," *Transportation Research Part B: Methodological*, vol. 34, no. 8, pp. 625–644, 2000.

[19] J. K. Eom, J. Y. Song, and D.-S. Moon, "Analysis of public transit service performance using transit smart card data in Seoul," *KSCE Journal of Civil Engineering*, vol. 19, no. 5, pp. 1530–1537, 2015.

[20] T. Stasko, B. Levine, and A. Reddy, "Time-expanded network model of train-level subway ridership flows using actual train movement data," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2540, no. 1, pp. 92–101, 2016.

[21] W. L. Wang, S. M. Lo, and S. B. Liu, "Aggregated metro trip patterns in urban area of hong kong: evidence from automatic fare collection records," *Journal of Urban Planning & Development*, vol. 141, no. 3, pp. 2692–2704, 2015.

[22] R. Xu, Y. Li, W. Zhu, and S. Li, "A multi-modal evacuation model for metro disruptions: based on automatic fare collection data in Shanghai, China," in *Proceedings of the 95nd Annual Meeting of the Transportation Research Board*, 2015.

[23] H. Niu, X. Zhou, and R. Gao, "Train scheduling for minimizing passenger waiting time with time-dependent demand and skip-stop patterns: nonlinear integer programming models with linear constraints," *Transportation Research Part B: Methodological*, vol. 76, pp. 117–135, 2015.

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

[25] M. H. Poon, S. C. Wong, and C. O. Tong, "A dynamic schedule-based model for congested transit networks," *Transportation Research Part B: Methodological*, vol. 38, no. 4, pp. 343–368, 2004.

[26] Y. Sun and P. M. Schonfeld, "Schedule-based rail transit path-choice estimation using automatic fare collection data," *Journal of Transportation Engineering*, vol. 142, no. 1, article 04015037, 2016.

[27] J. Shi, F. Zhou, W. Zhu, and R. Xu, "Estimation method of passenger route choice proportion in urban rail transit based on AFC data," *Journal of Southeast University (Natural Science Edition)*, no. 1, pp. 184–188, 2015.