Hindawi
Journal of Advanced Transportation
Volume 2023, Article ID 5596285, 19 pages
https://doi.org/10.1155/2023/5596285

Research Article

Passenger Flow Path Prediction Based on Urban Rail Transit AFC Data: An Example of Chengdu, China

Yu Wang,[1] Qixuan Qin,[1,2] Jialiang Chen,[3] Jiangbo Wang,[3] and Kai Liu[3]

[1]*School of Traffic and Transportation Engineering, Dalian Jiaotong University, Dalian 116024, China*
[2]*Department of Railway Traffic Operation Management, Baotou Railway Vocational and Technical College, Baotou 010010, China*
[3]*School of Transportation and Logistics, Dalian University of Technology, Dalian 116024, China*

Correspondence should be addressed to Qixuan Qin; 18147249266@163.com

Received 25 May 2023; Revised 21 July 2023; Accepted 12 October 2023; Published 10 November 2023

Academic Editor: Muqing Du

The development of the automatic fare collection (AFC) systems provides significant support for predicting passenger flow on urban rail transit. This paper extracts passenger travel patterns using AFC data on urban rail transit in Chengdu, China, over a one-month period. Passengers are divided into two categories based on their travel habits and data mining models, and multinomial logit (MNL) models are separately used to predict their destinations. Furthermore, a two-way search algorithm is developed to search the optimal paths between origin-destination (OD) pairs by considering interchange constraints. Start a path search through the origin point and destination point, respectively, until the shortest path is found. The maximum effectiveness of a path is measured by travel time, interchange time, and the number of interchanges between the OD pairs. Finally, the validity of the proposed passenger flow path prediction method is verified by using the AFC data of Chengdu metropolitan rail transit from April 2018.

1. Introduction

By March 2022, there were 49 cities in mainland China that had constructed urban rail lines, totaling 8,837 km, making it one of the fastest-growing countries in terms of urban rail transportation. As people's living standards continue to improve, higher demands are being placed on the safety, efficiency, and service levels of urban rail systems. The AFC System's ridership data provide information for passenger flow-related analysis and station/line status assessment. The accuracy of the model prediction results can be fully guaranteed by simulating and testing the established station passenger flow or passenger flow OD prediction model using historical AFC data. For example, Guo et al. [1], Tang et al. [2] predicted station passenger flows and validated them using historical AFC data; Yang et al. [3], Cao et al. [4], and Yao et al. [5] built an OD matrix prediction model and compared the prediction results with real data to verify their validity. However, people's travel patterns are heterogeneous [6] and may change over time. The models need to be updated with new indicators or calibration parameters due to changes in the operation of the urban transportation system (e.g., changes in urban land use types or the introduction of new routes). Using the previous model may lead to relatively inaccurate predictions.

To obtain relatively accurate prediction results of urban rail traffic, an increasing number of studies have used data mining on AFC data to extract information, such as station cross-sectional passenger flow and inbound and outbound passenger flow [7], or to obtain travel preferences of established cardholders to build prediction models with stronger generalization capability. Figure 1 shows the timeline of the search in the Web of Science core database with the search formula "(rail* or metro or subway or underground) and (forecast* or predict*) and passenger* and (AFC or OD)" (97 search results as of March 31, 2023), and the search results were imported into CiteSpace for visualization. In the two figures of Figure 1, time is

increasing in years along the time line from left to right, and the rows represent the category results of clustering, decreasing in number from top to bottom. The clustering results of Figure 1(a)) and Figure 1(b)) are consistent, but the difference is that Figure 1(a)) labels the literature with keywords and Figure 1(b)) is labeled with titles. Observing the two figures in Figure 1, most research on short-term passenger flow forecasting is followed by those considering forecasting methods in terms of spatiotemporal correlation, while #2 and #3 both forecast OD passenger flows and mostly use deep learning to provide algorithms for mining AFCs (more spanning lines between #2 and #3 implies more common literature in both categories). Moreover, in the last two years of research, passenger flow forecasting, especially OD passenger flow forecasting, has been studied further, revealing that there are indeed urgent problems in this area in the current period. Therefore, current research mainly tends to employ data mining algorithms for OD passenger flow prediction in urban rail transit, considering spatiotemporal correlation factors [1, 2, 8–13].

The urban rail transit OD passenger flow forecasting can be divided into two steps: D-point forecasting (also called OD matrix forecasting) and inter-OD flow allocation (also called inter-OD path selection). On the one hand, based on the passenger flow characteristics, passenger flow distribution patterns [14], and passenger travel preferences [7], combined with the urban rail topology network [15], it is possible to build a generalized model to measure the OD matrix of urban rail passenger flow for the prediction of passenger point of interest (POI) [16]. For example, the improved LSTM algorithm [15, 17, 18] is a more widely used method for predicting the OD matrix, and there are also nonlinear models [3], HW-DMD [19], etc. On the other hand, in the transportation domain, a trip is generally described by an OD pair, and there are usually many paths between each OD pair that can be chosen by the traveler. Initially, people may choose the path that costs the least amount of time, money, etc. to travel, i.e., the "shortest path." However, because of the combination of different factors such as the passenger's travel purpose, travel time [14], and the attractiveness of the destination station [20–22], passengers tend to choose the path with the least cost in a broad sense, which is called the path with the greatest effectiveness in transportation science. As the path with the greatest effectiveness is continuously chosen, an increasing number of people will be on this path, resulting in increased congestion and time costs, and the effectiveness values between the shortest path and the second shortest path will gradually approach, even if the shortest path can no longer be the shortest path. Therefore, path selection probability prediction and OD demand prediction are studied as branches of research on inter-OD traffic assignment. For example, some studies have predicted the paths chosen by groups by constructing probabilistic models of path selection [23] or by matching travel time clustering to OD routes [24]; others have predicted the OD demand by constructing improved LSTM models [4, 25], improved CNN models [26, 27], or for emergency [28] or COVID-19 periods [29].

In terms of data mining depth, current research is mainly divided into the extraction of overall indicators from AFC data, such as the direct extraction of inbound and outbound passenger flow and time-of-day passenger flow from AFC data [9, 30], or the mining of travel habits of specific passengers from AFC data and the use of set counting models to conduct research at the category level [4, 7, 22, 30, 31]. However, the current models cannot match the efficient response needs of real-time systems due to the sheer volume of their parameter systems, and the information obtained based on real-time AFC data is likely to be data that has not been fully populated due to data transmission lag and cannot be mined for historical travel preferences to obtain prediction results. The multinomial logit (MNL) model [32] is based on each passenger's choice and simulates the process of passengers deciding travel options. When passengers' travel habits are developed, the results of the choice will be closer to the actual situation because their travel perceptions will not change extensively in a short period, which is more suitable for station forecasting of passenger flow destinations in urban rail transit. Therefore, the logit model [33] and its improved form [34–37] are more interpretable and provide a more significant representation of passenger travel preferences and behaviors than the deep learning algorithm-based prediction approach described above.

In this study, we utilize a combination of data mining techniques and a logit model to predict passenger behavior for different passenger types. By analyzing massive historical automatic fare collection (AFC) data, we analyze the travel patterns of two distinct passenger groups - specific cardholders and those without prior travel data. In addition, we introduce area attractiveness to predict origin–destination (OD) matrix and identify effective travel routes. Our proposed method can be utilized for real-time passenger flow prediction in an online environment.

The paper is structured as follows. Section 2 provides a comprehensive description of the database used in the study. In Section 3, we construct the road network passenger flow OD dynamic estimation and passenger flow path assignment model. In Section 4, we demonstrate the numerical analysis approach to predict passenger flow paths and related issues. Finally, we summarize our research findings and propose future research directions in Section 5.

2. AFC Data of Network Passengers

China's rail transit system has basically implemented the automatic collection of passenger entry and exit information for AFC systems. We take an AFC dataset of Chengdu Metro Line 2 in China as an example to describe the structure of the AFC data, as shown in Table 1.

In the current situation of urban rail transit operation, AFC data usually have problems such as missing key information and abnormal data, resulting in poor data integrity and accuracy. To improve the accuracy of data mining, the "dirty data" in the historical AFC data should also be filtered, such as ticket card data lacking key information, data with duplicate records, data with identical OD points, illogical entry and exit times, data with numerous rides in a short period, or data with long travel times that do not conform to normal travel patterns.

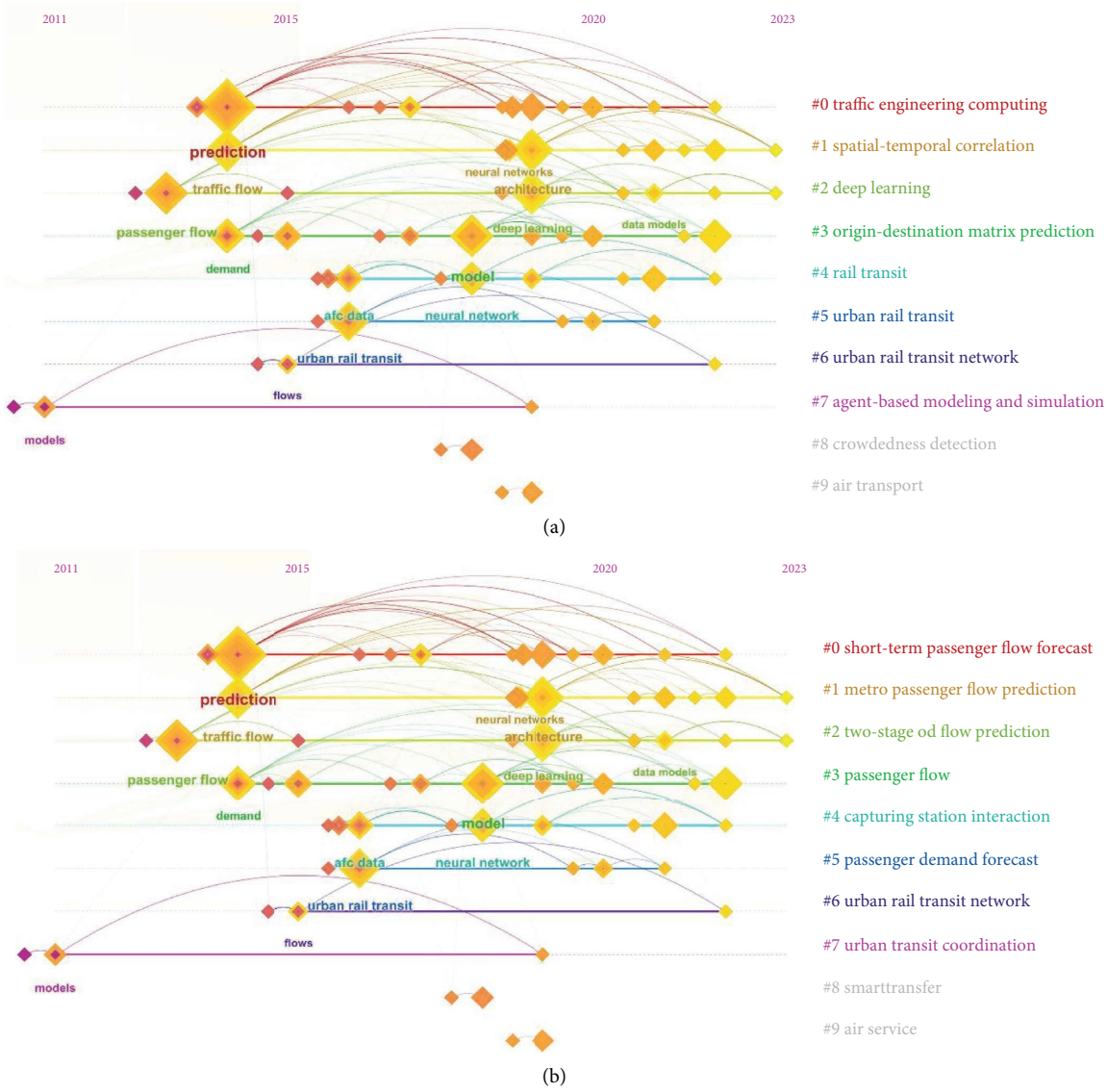

Figure 1: Visualization results of the literature in the field of rail AFC and OD. (a) Literature timeline results tagged with keywords. (b) Literature timeline results tagged with titles.

2.1. Site Type. To conduct an OD point analysis and identify rail stations with more intensive commuter traffic, it is essential to classify the stations. However, subdividing each of the 156 rail transit stations in Chengdu into multiple factors would require significant human resources, time, and effort. As an alternative approach, we classified Chengdu rail transit stations into seven distinct types based on the distribution of incoming and outgoing traffic over time. These classifications include: residential-concentrated, office-concentrated, residence-dominated residential-office, office-dominated office-residential, commercial-concentrated, hub, and other types, which can be found in Table 2.

From the above classification, it can be seen that the size of the inbound and outbound passenger flow of a station has a certain relationship with the attractiveness of the area around the station [22]. For example, the purpose of passenger trips in residential stations is mainly commuting to and from work, commuting to and from school, and shopping trips, while the purpose of passenger trips in office stations is mainly commuting to and from work. Therefore, different trip purposes also lead to different spatial and temporal distributions of OD between different sites.

3. Methodology

In this section, we construct an AFC-data-based passenger flow path prediction model for urban rail transit. As shown in Figure 2, the model consists of two parts: dynamic estimation of network passenger flow OD and passenger flow path assignment based on AFC data. Among them, the dynamic estimation of the network passenger flow OD model divides passenger travel data into two categories: travel habits and not-forming travel habits, and performs D-point prediction on the acquired urban rail transit route

Table 1: Partial sample of urban rail transit AFC swipe data.

ID card number	Card type	Inbound stations	Inbound date	Inbound time	Outbound stations	Outbound date	Outbound time
349xxxx	One-way tickets		2018/4/8	16:48		2018/4/8	17:31:20
61000001 1550xxxx	Stored-value tickets	Xipu	2018/4/8	16:40:58	Chengdu east passenger station	2018/4/8	17:31:27
61001000 7103xxxx	Regular CPU card		2018/4/8	16:44:59		2018/4/8	17:35:21

Table 2: Site type.

Station types	Station characteristics	Representative site
Residential-concentrated type	High flow of passengers entering the station in the morning peak and leaving the station in the evening peak during working days	Baiguolin
Office-concentrated type	High volume of outbound passengers in the morning peak and inbound passengers in the evening peak during working days	Gaoxin station
Residence dominated residential-office type	Two peak (the morning peak and evening peak), with the morning peak inbound traffic being larger than the evening peak inbound traffic and the opposite outbound traffic	North renmin road
Office dominated office- residential type	Two peaks (the morning peak and the evening peak); the morning peak outbound passenger flow is greater than the evening peak outbound passenger flow, while the inbound passenger flow is the opposite	Jinjiang hotel
Commercial-concentrated type	Include stations with scenic spots nearby. The flow of passengers is high throughout the day, and there is no obvious morning and evening peak traffic; on weekends and holidays, there are small peaks of traffic with a high degree of volatility	Chunxi road
Hub type	Includes stations with nearby railroad stations, passenger distribution points, etc., and stations at the intersection of urban rail lines No obvious pattern; the flow of passengers than other sites has always been larger; similar to the commercial concentration type, this type of site is not obvious with small peaks of passenger traffic, and generally larger passenger flow	Chengdu east passenger station
Other types	No obvious traffic pattern can be found, and the outbound traffic is not large	—

network according to the travel data categories. The passenger flow path assignment model determines the effective path set between the OD pairs by a two-way search algorithm and uses travel time, number of interchanges, and transfer time as the influencing factors, combined with the AFC data, to determine the final prediction results between the OD pairs.

3.1. Dynamic Estimation of Passenger OD Flow. To estimate the real-time passenger flow, a pattern analysis of passengers is needed to quickly find their chosen outbound station (point D) for all passengers entering the station at point O. However, in the AFC data of the urban rail passenger flow, the following four situations may occur, resulting in the unavailability of outbound station results:

(1) The amount of ridership data for a passenger is too small to reach the baseline value and is judged insufficient to form a travel habit.

(2) The base value is too high, resulting in the inability to filter out suitable outbound stations.

(3) The same entry information cannot be found in the history data of a passenger.

(4) The number of predicted D points in the output is greater than one.

We divide all AFC data into two categories: passengers who have formed travel habits and passengers who have not formed travel habits. The first three cases are grouped into the second group, and a group-level data mining strategy is carried out. For the fourth case, we narrow down the historical data matches by using count period segmentation to filter the similarity data in the time dimension.

The notations of the variables used in this section are given in Table 3.

3.1.1. Data Mining Algorithm. Among all historical records, it is clearly unreasonable to judge and classify passengers' travel habits by only one swipe of the card data. To determine the number of baseline values, we use the data of Chengdu metropolitan rail transit in April as sample data. Out of the total data, 5 million samples were taken, and all rides with the "Tianfutong Stored Value Ticket" (a long-term card held with high travel dependency) were selected and grouped by card number. In the judgment, the number of trips of the same ID card number is selected as the base number, i.e., if the base number is set to 1, all trip records of the same card number with the number of trips greater than 1 within the data are screened, and the amount of data conforming to the base number is output and specified as "Regular Number." In addition, the data were regressed by the name of the outbound station and compared to the last day of April to calculate the accuracy rate, and the results are shown in Table 4.

It can be seen from Table 4 that as the number of baseline values increases, the number of rides that conform to the regular number gradually decreases, but the accuracy rate gradually increases. When the baseline values are assumed to be three and four, the number of samples does not decrease significantly, but the accuracy rate increases substantially. To balance the constraints between the accuracy rate and the baseline value, we consider four as the baseline. In addition, the determined value of travel habits is calculated based on the actual situation of the Chengdu subway system ε, taking 35% of the experience value. The idea of the mining algorithm is as follows:

Step 1: Obtain the passenger entry information uploaded from the real-time AFC data and match the ID card number in the historical ride record database; if there is no matching result, deal with the data with the process of D-point prediction of passengers who have not formed travel habits (the method stated in Section 3.1.2)

Step 2: Count the number of rides corresponding to the ID card number and judge whether it is lower than the baseline value. If it is, we deal using the data with the method stated in Section 3.1.2. If it is greater than the baseline value, execute Step 3;

Step 3: Filter the passenger inbound station X_i corresponding to the current AFC data, and output the information of all outbound stations $X_1, \ldots, X_j, \ldots, X_n$ in the specified counting period (such as month, day, etc.), obtain $N_1, \ldots, N_j, \ldots, N_n$ by counting the number of times the passenger exits at each station, and calculate the travel habit determination value ε_j:

$$\varepsilon_j = \frac{N_j}{\sum_{j=1}^{n} N_j}. \tag{1}$$

Determine whether ε_j is greater than 35%, if it is greater than 35%, enter Step 4; if it is less than 35%, it is judged to be a passenger without travel habits, and the data of this ID card number are plugged into the method stated in Section 3.1.2

Step 4: Count the data whose value of ε_j is greater than 35%. If there are multiple data points, refine the counting period, then find the historical ride-out stations and return to execute Step 3. If there is only one data point, execute Step 5

Step 5: Output the outbound station corresponding to the value of ε_j, defined as the predicted outbound station X_j, which is the predicted outbound station for the passenger.

3.1.2. Spatiotemporal MNL Prediction Method Based on Unformed Travel Habits. According to the classification of passengers' travel habits, the travel data that do not reach above the baseline value indicate that the passengers corresponding to such data have not yet been explored for travel habits and cannot be pinpointed. For passengers who have not yet formed travel habits, we treat them as a group and perform a group probability distribution study because we cannot analyze the historical preference data of individual passengers.

The probability function $P_p(j)$ of the MNL model describes the probability that a choice set j (in our study, the choice set denotes the outbound station chosen by passenger p) will be chosen.

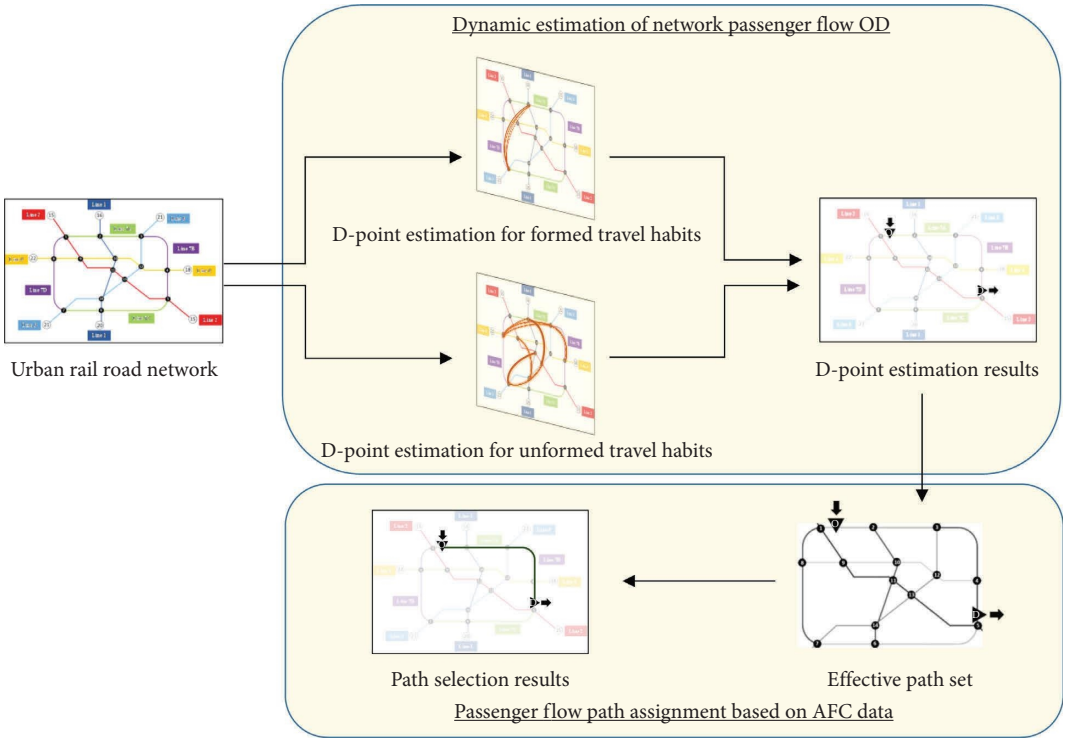

Figure 2: Flow diagram of the method in this chapter.

Table 3: Notation and description of the OD dynamic estimation model.

Notation	Description
Baseline	In the historical period contained in the AFC data, when the number of times a passenger enters the station with the same ID card number is greater than or equal to four. It is determined that the passenger has reached the baseline value and has a certain travel pattern
p	The number of the passenger corresponding to the ID card number
i	An inbound station. $i = 1, 2, \ldots, m$.
j	When the inbound station i is determined, the corresponding passenger outbound station j. $j = 1, 2, \ldots, n$
N_j	The number of exits of passengers from station j during the counting period
ε	When a passenger corresponding to a certain ID card number has a record of an inbound swipe that reaches the baseline value at the same station, the number of times that a certain outbound station j accounts for a percentage greater than or equal to ε_i of all outbound stations corresponding to that inbound station, the D-point prediction result for that passenger is j
$P_p(j)$	The probability that passenger p chooses outbound station j when travelling
U_{ij}	The effectiveness function of passenger p travels when i enters - j leaves. The effectiveness function is expressed quantitatively in terms of the factors affecting the choice of the outbound station. The effectiveness function is divided into a fixed term and a probability change term
C_{jp}	A fixed term of the effectiveness function U_{ij} for passenger p to select the outbound station j
A_p	Passenger p travels to select all choice sets of the outbound station j. The number of elements of the choice set A_p determines the number of items in the logit model
T_{ij}	Total travel time between inbound and outbound stations (in seconds) [37]. In the AFC data without clear travel habits, the short distance between ODs is associated with a high probability of being selected and high traffic volume, so it is assumed that the higher the traffic volume between stations, the shorter the travel time, i.e., there is a negative relationship between the traffic volume between stations and travel time

Table 3: Continued.

Notation	Description
α	Coefficients of total travel time variables
D_j	Outbound station regional attractiveness, describing the number of people (in person) that the outbound station can attract [19]. Define the regional attractiveness as 0-1 variable, and the outgoing number of outbound stations is judged to be attractive and defined as 1 if it is greater than the average value; it is defined as 0 if it is less than the average value
β	Coefficients of regional attractiveness variables
G_{ij}	The OD size variable, noted by rank, describes the effect of inbound and outbound station size on OD volume. The average value of inbound and outbound traffic corresponds to rank 5, and the maximum inbound and outbound traffic corresponds to station rank 10
γ	Coefficients of origin and destination scale variables

Table 4: Judgment of the relationship between the base number, the regular number, and the accuracy rate.

Baseline value	Regular number	Accuracy rate
1	969073	0.834
2	660626	0.898
3	533521	0.918
4	420108	0.946
5	344493	0.957
8	154567	0.963
10	29883	0.974
15	2038	0.979
20	181	0.984

$$P_p(j) = \frac{e^{C_{jp}}}{\sum_{j \in Ap} e^{C_{jp}}}. \quad (2)$$

In general, the greater the attractiveness of the area, the shorter the travel time, and the greater the volume of passengers, the greater the probability of passengers exiting the station. Therefore, we define the effectiveness function U_{ij} as a linear function based on the MNL model with the following mathematical expression:

$$U_{ij} = \frac{\alpha T_{ij}}{3600} + \beta D_j + \gamma G_{ij}. \quad (3)$$

All U_{ij} in the above equation are fixed terms, so that $C_{jp} = U_{ij}$ in the probability function. Therefore, the spatiotemporal MNL model is constructed jointly with equations (5) and (6) to predict the outbound station (D) corresponding to a given inbound station (O).

Using the April 2018 Chengdu city rail transit data as sample data for regression analysis, it is possible to determine the parameter estimates generated by each factor affecting the effectiveness function on the choice of passenger D points, thus calibrating the effectiveness function equation (6) parameters of the D point prediction logit model.

(1) Travel Time. T_{ij}Clustering. The travel time of T_{ij} from Gaoxin station to each station is selected as the case, and because of the variability of individual samples, the average travel time is considered the value of T_{ij}. Meanwhile, stations with smaller outbound passenger flow will be filtered out, and the travel time from Gaoxin station to each station is finally estimated, as shown in Figure 3.

(2) Regional Attractiveness. D_jQuantified. To facilitate the analysis, we will select the most representative station of each type, count its interstation traffic in April, and then compare the average value of the inter-OD traffic of all stations. If the actual traffic is greater than the average value, the result is recorded as 1; otherwise, it is recorded as 0. The results of the inter-OD traffic calculation for the six representative stations are shown in Table 5.

For the statistical classification of the OD passenger flow between different types of sites, the OD volume between different types of sites is compared with the average value of the OD volume of 12,297 passengers of all sites, and if the actual OD volume is greater than the average value, it is recorded as 1; otherwise, it is recorded as 0. The results are shown in Table 6.

(3) Quantification of Scale Variable. G_{ij}. In the AFC data of Chengdu city rail transit used in the study, the average value of inbound and outbound station traffic for all stations was calculated as 884,565, and the maximum inbound and outbound station traffic was 6,165,033 at Chunxi Road. According to the grade progression, the grade increases by one for each million increase in traffic after grade 5; before grade 5, the grade increases by one for each 170,000 increase in traffic. Thus, the total inbound and outbound passenger flow at the major stations of the Chengdu Metro from April 1 to 30 is shown in Figure 4.

Based on the maximum likelihood method for estimation [38], the values of each parameter of the calibrated effectiveness function equation (6) are taken, and the results are $\alpha = -0.2310$, $\beta = 0.2132$, and $\gamma = 0.3387$.

3.2. Passenger Flow Path Assignment Based on the AFC Data. In this section, the effective path topology model is first established and solved to obtain the effective path set, and then the path selection model is used to calculate the selection probabilities of different paths to realize the refined passenger flow allocation.

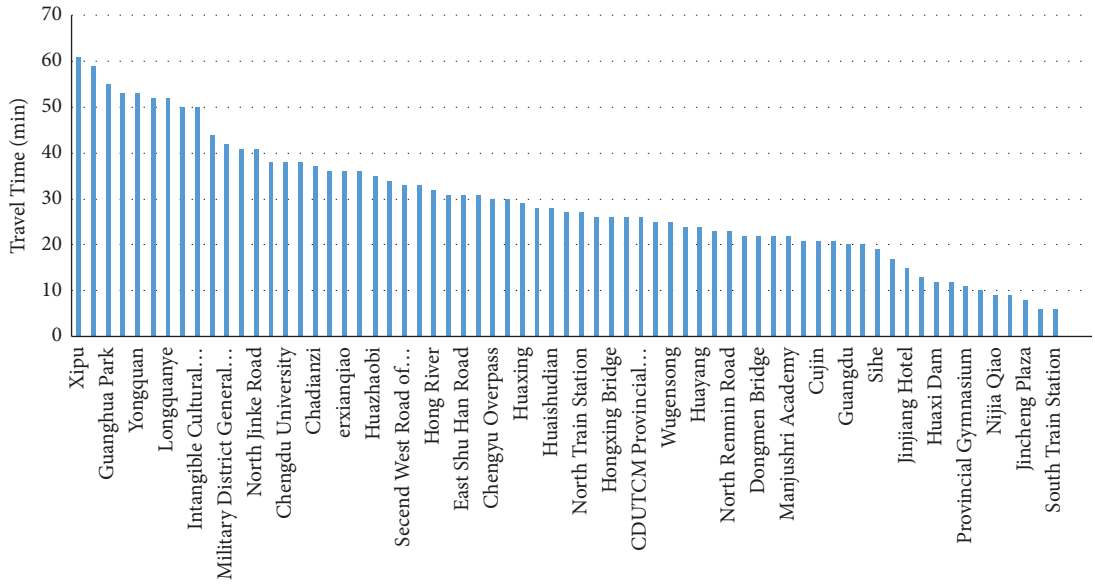

Figure 3: Average travel time between Gaoxin station (O) and other stations (D).

Table 5: OD traffic at representative sites.

O\D	Chengdu east passenger station	Jinjiang hotel	Gaoxin	North renmin road	Baiguolin
Chunxi road	136183	13064	47918	56875	55495
Baiguolin	34771	3254	9989	5637	—
Gaoxin	30169	23292	—	31086	—
North renmin road	41795	12421	—	—	—
Jinjiang hotel	17245	—	—	—	—

Unit: person.

Table 6: OD traffic distribution and site type.

O\D	Office-concentrated type	Residential-office type	Office-residential type	Commercial-concentrated type	Hub type
Residential-concentrated type	0	0	0	1	1
Office-concentrated type	—	0	1	1	1
Residential-office type	—	—	1	1	1
Office- residential type	—	—	—	1	1
Commercial-concentrated type	—	—	—	—	1

The symbols of the models covered in this section and their interpretations are shown in Table 7.

3.2.1. Effective Path Topology Model and Solving Algorithm. Considering that the algorithm needs to conform to the actual travel habits of passengers travelling normally, the following assumptions are made:

(1) Stations and each interval section of the urban railway can only be passed once.

(2) The number of interchanges for passengers using urban rail transit is limited, i.e., the number of interchange stations in the effective path is limited. According to experience, the number of interchanges from the original point to the destination station is generally not more than three.

(3) Passengers who use urban rail transit to travel will generally not transfer into a line again if they change out of a line when transferring. That is, the paths in the effective route are continuous on each rail line.

(4) If the passenger's OD point is on the same route, the passenger will only travel on that route, i.e., if the OD point is on the same route, the valid trail is also on the same route (hypothesis 3 and hypothesis 4 are complementary to each other).

In turn, the rail network is transformed into a directed connectivity graph $G = <V, E, T>$ to describe the rail network

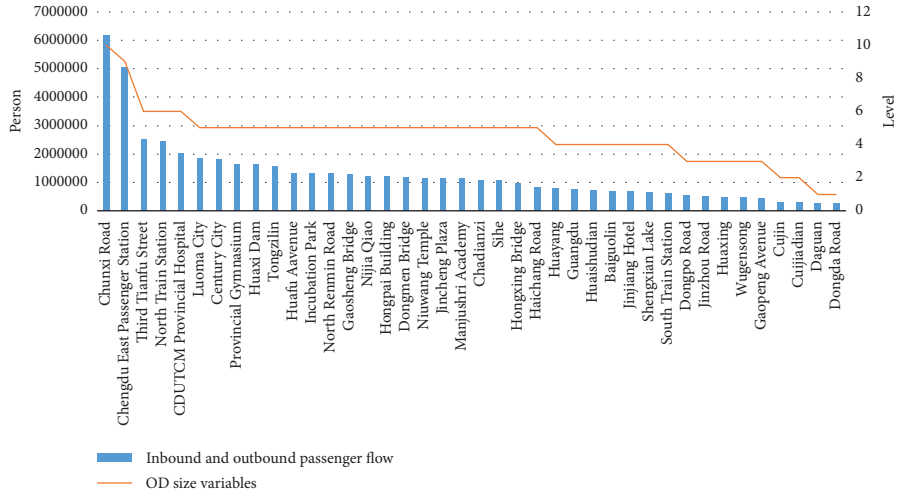

Figure 4: Passenger flow and OD size of each station.

Table 7: Passenger flow distribution model notation and description.

Notation	Description
v_k^l	The station with the number k and the line l to which it belongs; if v_k^l is an interchange, then k is the number of the interchange online l
$e_{k_1,k_2}^{h,d}$	The interval with the original station number k_1, the ending station number k_2, the direction d and the number h; $d = ((k_1 - k_2)+1/2)$
$t_{k_1,k_2}^{m,(d_1,d_2)}$	The virtual interchange section with the starting interchange level number k_1, the ending interchange level number k_2, the interchange direction (d_1, d_2), and the number m is connected
W	Effective path set
T_i^O	Adjacent stations of the original station O, including $T_1^O T_2^O$
T_i^D	Adjacent stations of the ending station D, including T_1^D, T_2^D
L_l	The collection of stations (including interchange stations and ordinary stations) included in the urban rail line of l th
D	Line interval collection
S	Collection of stations along the valid path
B	Effective path along the interval collection
n	Combined count parameter of adjacent stations, i.e., number of interchanges
m	Valid path count parameters
K_{rs}	Set of valid paths between OD pairs
U_k^{rs}	The random effectiveness of the traveler's choice of an efficient path k between OD pairs, $k \in K_{\text{rs}}$
V_k^{rs}	The determined effectiveness value of the traveler can be expressed in terms of the path cost
$\varepsilon_k^{\text{rs}}$	Random error
θ	Constant, inversely proportional to the variance of $\varepsilon_k^{\text{rs}}$, is commonly used as a conversion of route cost into effectiveness. It can be interpreted as an indicator of the overall familiarity of passengers with the urban rail network
p_k^{rs}	Probability of choosing a valid path k for $r - s$ passengers between OD pairs
C_k^{rs}	The subset of valid paths of a valid path k, the smallest set of selectable paths for passengers
$C_{\min}^{\text{rs}}$	The shortest path cost between OD pair $r - s$
T_{tra}	Passenger transfer time
t_{twa}^i	The travel time (s) of the i th transfer, i.e., the time from the disembarkation platform to the waiting platform for the passenger's i th transfer
t_{tpl}^i	Waiting time at the platform for the i th transfer (s)
t_{L_l}	Passenger travel time online l, i.e., the transfer arc consumption time for a particular transfer, obtained from the survey
f_l	The interval between departures of line l in a particular interchange
T_{rs}^k	Travel time on the path k
n_k	Number of interchanges on the path k
T_{tra}^k	Interchange time on path k
H	Nonnegative constant, called the stretch factor of the path

model by hierarchical sequencing of the network, where V is the set of stations, E is the set of intervals, and T is the set of interchange virtual intervals [38, 39].

$$v_k^l \subset V, e_{k_1,k_2}^{h,d} \subset E, t_{k_1,k_2}^{m,(d_1,d_2)} \subset T. \tag{4}$$

Given any OD points, let the set of ordered intervals contained in the valid path be C, where the actual interval ordered set is X, and the virtual interval ordered set is Y. Then, the valid path should satisfy the following conditions:

$$\begin{aligned} C = \{ & \cdots, x^{j-1}, x^j, y^{j+1}, \\ & x^{j+2}, x^{j+3}, \cdots \}, \end{aligned} \tag{5}$$

$$X = \{\cdots, x_1^i, x_1^{i+1}, \cdots\}, \tag{6}$$

$$Y = \{\cdots, y^n, y^{n+1}, \cdots\}, \tag{7}$$

$$X + Y = C, \tag{8}$$

$$k_2^{x^j} = k_1^{x^{j+1}}, \tag{9}$$

$$k_2^{x^j} = k_1^{y^{j+1}}, \tag{10}$$

$$k_2^{y^j} = k_1^{x^{j+1}}, \tag{11}$$

$$l_2^{y^{n_1}} \neq l_1^{y^{n_2}}, \tag{12}$$

$$\sum y \leq N, \tag{13}$$

$$\begin{aligned} & i = 1, 2, 3, \cdots \\ & j = 1, 2, 3, \cdots n = 1, 2, 3, \end{aligned} \tag{14}$$

where l_1 and l_2 denote the lines belonging to the virtual intervals connecting the OD stations k_1 and k_2, respectively. Equation (12) represents the two real intervals x^j and x^{j+1} adjacent to each other in the ordered set C. The k_2 of the former is the same as the k_1 of the latter to ensure the continuity of the effective paths on the same line. Similarly, equations (13) and (14) ensure the continuity between the actual and virtual intervals of the effective paths during the interchange process. Equation (15) represents any two different virtual commutation intervals y^{n_1} and y^{n_2}, in the effective path, with l_2 of the former being different from l_1 of the latter, so that the effective path satisfies the basic assumption (3). Equation (16), limiting the number of interchanges n satisfies the basic assumption (2).

To reduce the complexity of the algorithm and solve the above effective path topology model, we store the road network information in the station number in advance, omit the step of introducing the adjacency matrix, reasonably use the feature that the number of interchanges n does not exceed 3 times, and adopt the "two-way search algorithm" with both O and D points as the starting points as the effective path set solving algorithm. The steps of the "two-way search algorithm" are as follows:

Step 1: Initialize the effective path set W, original station O, destination station D, $L_l = \{v_1, v_2, \ldots, v_k\}$ (k is the total number of stations), D, S, B, n, and m.

Step 2: Determine the adjacent interchange stations T_1^O, T_2^O, T_1^D, and T_2^D according to the original station O and the destination station D. If one end is the end station or is itself an interchange station, only one adjacent interchange station needs to be determined. Based on the line where the two adjacent stations are located, determine the line where the station is located for comparison. Then, determine whether the OD points are on the same line (based on our line number); if yes, then go to Step 5; if not, then go to Step 3.

Step 3: Cross-determine whether adjacent interchange stations are on the same line, discriminate up to four groups in total: (T_1^O, T_1^D), (T_1^O, T_2^D), (T_2^O, T_1^D), and (T_2^O, T_2^D), and denote their order by n. Initialize $n = 1$. If on same line, a valid path is found. For example, if (T_1^O, T_1^D) is discriminated on the same line, then a valid path $O \longrightarrow T_1^O \longrightarrow T_1^D \longrightarrow D$ expressed by interchange can be determined, and the path is stored in the set of valid paths W. Meanwhile, $n = n + 1$. If not on a line, let $n = n + 1$. When $n > 4$, go to Step 4.

Step 4: Since the algorithm specifies that the maximum number of interchanges n is three, when adjacent interchange stations are not on the same line with each other, to determine a valid path, one must find a station that satisfies the following requirements: the station is simultaneously on the same line with one of the adjacent stations at point O and on the same line with one of the adjacent stations at station D. Therefore, search for each of the four groups of adjacent stations, reinitialize $n = 1$, starting from (T_1^O, T_1^D): search for line L_1 where station T_1^O is located, search for line L_2 where station T_1^D is located, and then search for stations that belong to both L_1 and L_2, i.e., interchange stations of the two lines. If T_t exists, a valid path $O \longrightarrow T_1^O \longrightarrow T_t \longrightarrow T_1^D \longrightarrow D$ represented by interchange stations can be determined, and the path is stored in the set of valid paths W while $n = n + 1$; if it does not exist, let $n = n + 1$. When $n > 4$, go to Step 5.

Step 5: Initialize $m = 1$, starting from the first path in the set of valid paths W and determine each station and interval passed along the way. First, determine the specific route from the original point O to the adjacent interchange stations in the valid path, which shall be marked by two stations, determine the up and down direction, and retrieve the stations between the two stations together with the two stations deposited in the station set S. Second, retrieve the stations between the next two interchange stations by the same method until the end point is reached, and store the stations in the station set S. Third, by the order of stations in the station set S, according to the line interval set D, retrieve the square and conforming interval numbers between two stations in turn and deposit them in the interval set B. Let $m = m + 1$, and when $m > n$, go to Step 6. Otherwise, repeat the above steps.

Step 6: There are still many invalid paths obtained by the above algorithm because they do not satisfy the assumption (1) that a station or interval can be passed only once. Therefore, initialize $m = 1$ again and determine whether there are duplicate items in the set S of stations and the set B of intervals. If there is, delete the i th path in the set of valid paths W (where $i = m$). Let $m = m + 1$; when $m > n$, go to Step 7. Otherwise, repeat the above steps.

Step 7: Output the final set of valid paths W, and the algorithm ends.

3.2.2. Path Selection Model and Path Effectiveness Function. The passenger flow assignment problem is also commonly described as a path matching problem, where the probability of a passenger choosing a particular path reflects the degree of matching between the passenger flow and the path and can be expressed as the percentage of passengers choosing this path among all passengers. On the other hand, in transportation field research, the effectiveness function generally refers to the broad cost of a certain transportation mode or a certain path, which represents the functional relationship between the travel impedance perceived by the traveler and the travel influencing factors. Therefore, the path selection model is constructed with the basic logit formula as follows:

$$U_k^{\mathrm{rs}} = V_k^{\mathrm{rs}} + \varepsilon_k^{\mathrm{rs}}\, k \in K_{\mathrm{rs}}, \tag{15}$$

$$V_k^{\mathrm{rs}} = -\theta C_k^{\mathrm{rs}}\, k \in K_{\mathrm{rs}}, \tag{16}$$

$$p_k^{\mathrm{rs}} = \Pr\left(U_k^{\mathrm{rs}} \geq U_n^{\mathrm{rs}}, n \neq k\right) k \in K_{\mathrm{rs}}. \tag{17}$$

Clearly, the selection probability has the following properties:

$$0 \leq p_k^{\mathrm{rs}} \leq 1 \text{且} \sum_{k \in K_{rs}} p_k^{\mathrm{rs}} = 1, k \in K_{\mathrm{rs}}. \tag{18}$$

The probability p_k^{rs} of path selection is related to the distribution of the random error term $\varepsilon_k^{\mathrm{rs}}$ and the path cost C_k^{rs}. To reduce the irrationality of network traffic distribution, the relative cost can be used to calculate the selection probability. Assuming that the $\varepsilon_k^{\mathrm{rs}}$ are independent of each other and obey the Gumbel distribution, the path selection probability p_k^{rs} can be expressed in the following logit form by substituting the above equation into equation (5) in Section 3.1.2:

$$p_k^{\mathrm{rs}} = \frac{\exp\left(-\theta C_k^{\mathrm{rs}}/C_{\min}^{\mathrm{rs}}\right)}{\sum_k \exp\left(-\theta C_k^{\mathrm{rs}}/C_{\min}^{\mathrm{rs}}\right)}\, k \in K_{\mathrm{rs}}. \tag{19}$$

Based on the travel characteristics of urban rail transit, the main influencing factors considered by passengers in the process of path perception and selection are travel time, ride time, and the number of transfers. Since the current urban rail transit control system basically achieves a certain control accuracy and can ensure that the trains run according to the interval running map and train schedule, the ride time is regarded as a fixed constant that can be obtained from the train running map or train schedule. Therefore, the fixed term of random effectiveness in the path effectiveness function can be measured by three indicators: travel time, transfer time, and the number of transfers.

(1) Calculation of Travel Time. T_{rs}. We estimate the passenger flow distribution of multiple paths by determining the single-path passenger flow distribution. We used the travel time and number of passengers from 8:00 am to 10:00 am from April 10 to April 12 as the data samples for the Chengdu subway station "Xipu" to "Chunxi Road" and determined the travel time distribution function. The parameters of the distribution function were defined. The length of the interval was set at 30 s, and the statistical results are shown in Figure 5, using the "98th percentile" theory to eliminate the extreme minima at both ends.

Using hypothesis testing and the great likelihood estimation method, we determine that the travel time of the single path OD obeys a log-normal distribution $\ln N(\mu, \sigma)$ within the interval $[t_{\min}, t_{\max}]$ and has a parameter value $\hat{\mu} = 7.76617, \hat{\sigma} = 0.03623$. Thus, the mathematical expectation of the travel time from Xipu to Chunxi Road is $E(X) = e^{\mu+\sigma^2/2} = 2361\ s$.

The travel time probability distribution of multipath OD is the accumulation of different parameters of the normal distribution. Taking the OD from Xipu to South Railway Station as an example, there are two valid paths between this OD point pair. We establish a system of quadratic equations by using the data between the extreme value points of the frequency of Figure 6 (the red bar graph in the figure) as the data for the calculation of the system of equations, solving for the parameters of the normal distribution of the two paths, and solving for $\hat{\mu}_1 = 7.90$, $\hat{\sigma}_1 = 0.068$, $\hat{\mu}_2 = 7.95$, $\hat{\sigma}_2 = 0.046$. Therefore, the travel time expectation of path 1 is obtained: $E(X_1) = e^{\mu_1+{\sigma_1}^2/2} = 2839s$; the travel time expectation of path 2: $E(X_2) = e^{\mu_2+{\sigma_2}^2/2} = 2704\ s$.

(2) Calculation of Interchange Time. T_{tra}. Since the moment that passenger arrival at the platform is totally random, it is assumed that the arrival of passengers follows a uniform distribution over the interval $[0, f_l]$. Thus, the mathematical expectation of the passenger transfer waiting time at the platform is $0.5 f_l$.

Then, the interchange time calculation formula can be expressed as follows:

$$T_{tra} = \sum_{i=1}\left(t_{\mathrm{twa}}^i + t_{\mathrm{tpl}}^i\right) = \sum\left(t_{L_l} + 0.5 f_l\right) \forall l, i. \tag{20}$$

(3) Calculation of the Number of Interchanges. n. The number of interchanges n can be determined directly from the calculation results of the effective path search algorithm, and the algorithm is not described here. Note that if the path contains a virtual interchange arc, the interchange station is not counted in the number of interchanges.

In summary, the effectiveness function of path k for the broad cost, measured in terms of travel time, transfer time, and number of transfers, is as follows:

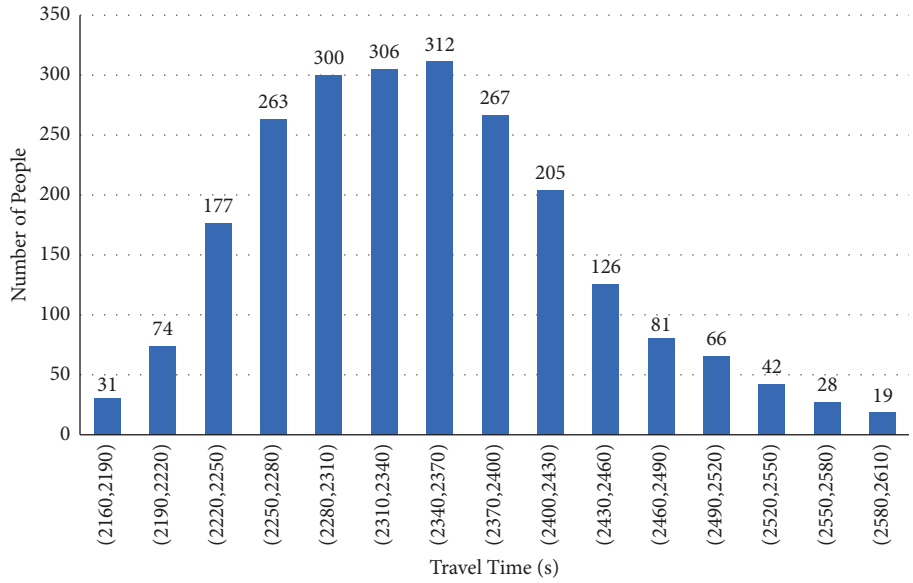

Figure 5: Histogram of "travel time-number of people" on Xipu-Chunxi road.

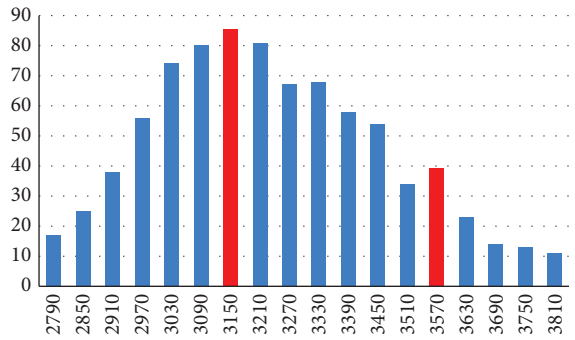

Figure 6: Histogram of "travel time-number of people" statistics of Xipu to south railway station red bar graph in the figure means the extreme value points of the frequency.

$$C_k^{\mathrm{rs}} = T_{\mathrm{rs}}^k + \alpha \cdot (n_k)^{\beta} \cdot T_{\mathrm{tra}}^k. \tag{21}$$

Equations (19) and (21) together form the route selection model where $\alpha \cdot (n_k)^{\beta} \cdot T_{\mathrm{tra}}^k$ is the interchange cost and α and β are parameters to be determined. Since the negative perception of passengers increases exponentially with each increase in the number of interchanges, β is an exponential parameter.

When there are multiple paths between ODs, it is necessary to study the selection behavior of passengers based on the elements of the path set. When passengers choose travel paths, they usually do not stand on the road network to consider all paths but choose from a part of the paths. Although we search the effective path set by a two-way search algorithm, the path set still contains too many paths, and in the actual passenger selection, one to three paths usually reach the limit. To find a subset of the valid path set, a stretch factor H is attached to all paths. Then, the subset of valid paths satisfies the following conditions [39, 40]:

$$C_k^{rs} \le (1 + H) C_{\min}^{\mathrm{rs}} \, k \in K_{rs}. \tag{22}$$

Therefore, after substituting the effectiveness function into the path selection function, the path selection model has four pending parameters: H, θ, α, and β. Using Chengdu City's data for calibration, we obtain $\alpha = 1.2720$, $\beta = 1.8623$, $H = 0.25$, and $\theta = 1.840$.

4. Results and Discussion

Since our model serves to judge the distribution of commuter traffic within the rail network during the commuter peak period, the data used in this section should select a station with high commuter traffic and an incoming passenger flow of ten minutes during the commuter peak period. Therefore, we chose all incoming swipe information from the Gaoxin station during 8:20 am–8:30 am on April 9, 2018, as the simulated real-time AFC upload data. In addition, to facilitate the observation of the regularity of the data, we selected the first four types of stations with a high number of outgoing stations in Table 2.

4.1. Outbound Station Prediction for Type I Passengers Based on Historical Travel Habits. During the period of 8:20 am–8:30 am on April 9, 2018, there were 99 swipe card data points entering the station at Gaoxin Station, distinguished by the ID card number of the incoming swipe card, indicating that 99 passengers entered the station. After filtering out the unrecorded card data and filtering out the passengers with a travel factor greater than 4, the remaining records are 67. Due to space limitations, we could not spread all ridership information here, so we chose two of the ridership data, as representatives to compare the results.

Passenger A and passenger B have historical AFC records, as shown in Table 8. Passenger A made 69 trips in a month, including 29 trips at Gaoxin Station; Passenger B made 49 trips in April, including 25 trips at Gaoxin Station.

Table 8: Record of the outbound records of a passenger after entering the Gaoxin station.

Passenger number	Time	D-point name	Number of outbound stops at D-point	ε_i (%)
Passenger A	April 2018	People's park	19	65.52
		Gaoxin	1	3.45
		Dongmen bridge	9	31.03
Passenger B	April 2018	Gaopeng avenue	13	52
		Hongxing bridge	11	44
		Chunxi road	1	4
Passenger B	April 2018 8:00–9:00 am	Gaopeng avenue	12	80
		Hongxing bridge	2	13.33
		Chunxi road	1	6.67

Calculate ε_i for the two passengers at their respective outbound stations in April, as shown in Table 8. For Passenger A, since the only station with $\varepsilon_i > 35\%$ is People's Park, Passenger A is predicted to leave the station at People's Park. For passenger B, since ε_i >35% corresponds to the two stations of Gaopeng Avenue and Hongxing Bridge, the time in the historical AFC data should be subdivided again, and all historical ridership data from 8:00 to 9:00 a.m. In the entry time of this card number history should be filtered, as shown in Table 8, and passenger B's travel habit determination value ε_i should be calculated again, and the one that exceeds 35% is Gaopeng Avenue, so passenger B's predicted outbound station is Gaopeng Avenue.

The real ridership records of passenger A and passenger B on the day of April 9, 2018, are shown in. From the exit information in Table 9, the outbound station that this predicted passenger would choose is the same as the actual outbound station.

Calculating the predicted results of passengers' outbound stations in that period corresponding to the above 67 data points, there are only 4 data points whose predicted stations do not match with the actual stations selected by passengers, and the prediction accuracy rate $\lambda = 94.03\%$. This mining algorithm is more accurate and reliable in calculating passenger outbound station selection for commuter flow.

4.2. Outbound Station Prediction for II Passengers Based on a Spatiotemporal MNL Model. For passengers who have not yet formed a travel habit, the examples are mainly passengers with one-way tickets and passengers with Tianfutong stored value tickets, Tianfutong cash cards, and Tianfutong regular CPU cards with less than four total trips in the historical AFC data.

Substituting the parameter values α, β, γ into the effectiveness function (6) of the spatiotemporal ML model is expressed as:

$$U_{ij} = \left(\frac{-0.421 \times T_{ij}}{3600}\right) + 0.2132 \times D_j + 0.2217 \times G_{ij}. \quad (23)$$

According to the travel time from the Gaoxin station to each station in Figure 3, we can see from equation (26) that the effectiveness function U_{ij} has a negative relationship with the travel time T_{ij} of passengers, so when T_{ij} is larger, the probability of passengers choosing the station is smaller, so the stations with travel time $T_{ij} \geq 30$ min are screened out first because their travel time is too long, so the probability of passengers choosing the station will be greatly reduced. In addition, when the travel time between two stations is too short, the possibility of passengers choosing other travel modes, including bus, walking, or bike-sharing, increases greatly, thus filtering out stations with travel time $T_{ij} \leq 10$ min.

Meanwhile, referring to Tables 2 and 6, the Gaoxin station is an office-concentrated station, thus calculating the $P_n(i)$ and U_{ij} of each outbound station corresponding to the Gaoxin station, as shown in Figure 7:

From the Figure 7, we can see that if we use the Gaoxin station as the inbound station for prediction, the vast majority of passengers will choose the station with a larger probability value $P_n(i)$ and effectiveness function U_{ij} as the outbound station, i.e., the station with the largest number of outbound passengers in this example should be Chunxi Road, Chengdu East Passenger Station, North Train Station, Tird Tianfu Street, and Provincial Stadium.

Extract the real card entry information for the corresponding date of Gaoxin station, screen out the stations with fewer than 90 exiters, sort the outbound stations according to the probability distribution, and obtain the following: Figure 8. Figure 8 shows that Chunxi Road, Chengdu east passenger station and north train station have the highest number of exits, with third Tianfu street and the provincial stadium ranking slightly differently. However, the change in traffic between third Tianfu street and the provincial stadium is not very different, so overall, the forecast results are more in line with expectations.

From the statistics, we can see that after a certain period of time, Chunxi Road, Chengdu east passenger station, and north train station will usher in a small peak of passenger flow, and the staff at these stations can deploy and plan the route of passengers in advance and conduct passenger flow diversion work at the right time to help the passenger flow evacuate quickly and avoid the formation of congestion.

4.3. Passenger Final Route Prediction. Most of the OD points in the preceding example are on the same urban rail line, and the distance is relatively short, which is not enough to illustrate the problem of multiple path selection. We reselect the "Chadianzi" station of Line 7 as the O point and the "Yinghui Road" station of Line 7 as the D point as the path prediction example in this section, as shown in the black inverted triangle in Figure 9.

Table 9: Actual card swipe record of passengers.

Passenger number	Card type	Inbound station name	Inbound time	Outbound station name	Outbound time
Passenger A	Tianfutongstored-value tickets	Gaoxin	"08:24:37"	People's park	"08:43:03"
Passenger B	Tianfutongstored-value tickets	Gaoxin	"08:21:46"	Gaopeng avenue	"08:33:57"

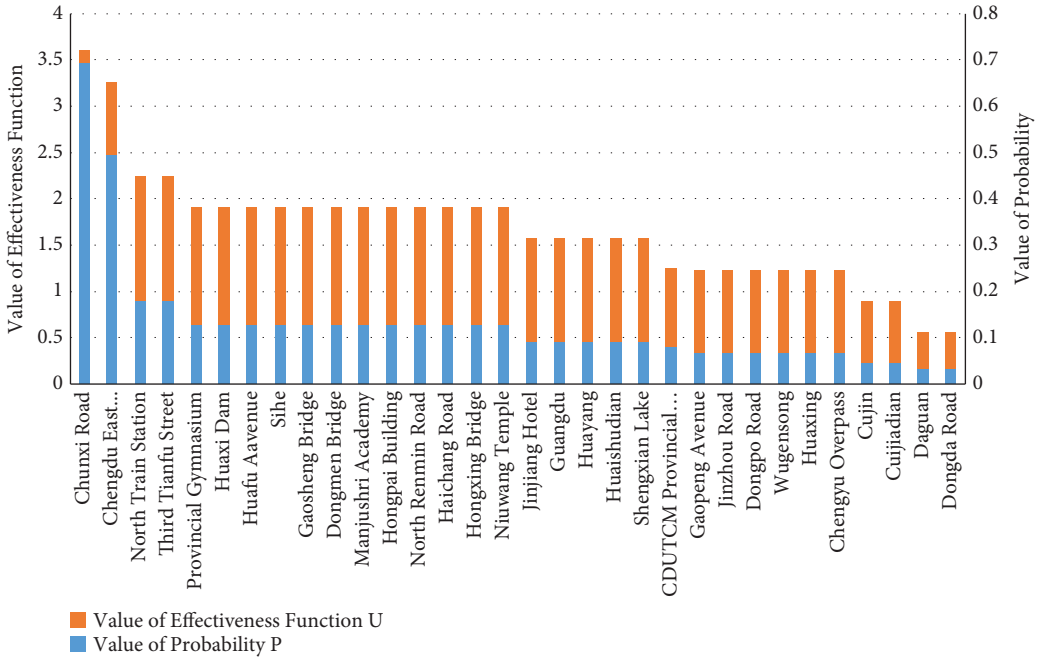

Figure 7: Outbound site probability and effectiveness function prediction results.

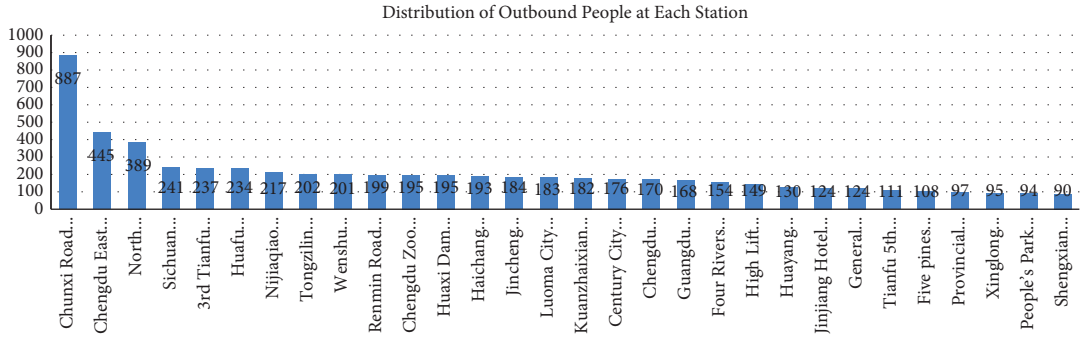

Figure 8: Distribution of the actual card, swiping in and out of the station.

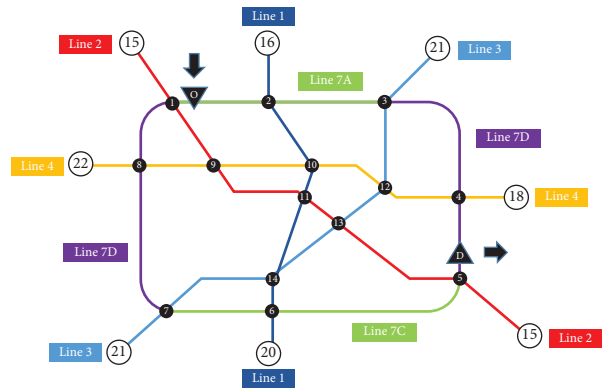

Figure 9: Chengdu rail transit network.

Table 10: Effective path search process with three interchange stations.

Adjacent interchange stations	Line transfer mode	Third interchange station	Effective path
① to ④	Line 2 to line 7B	⑤	O ⟶ ① ⟶ ⑤ ⟶ ④ ⟶ D
	Line 2 to line 4	⑨	O ⟶ ① ⟶ ⑨ ⟶ ④ ⟶ D
	Line 7A to line 7B	③	O ⟶ ① ⟶ ③ ⟶ ④ ⟶ D
	Line 7A to line 4	None	—
	Line 7D to line 7B	None	—
	Line 7D to line 4	⑧	O ⟶ ① ⟶ ⑧ ⟶ ④ ⟶ D
② to ④	Line 1 to line 7B	None	—
	Line 1 to line 4	⑩	O ⟶ ② ⟶ ⑩ ⟶ ④ ⟶ D
	Line 7A to line 7B	③	O ⟶ ② ⟶ ③ ⟶ ④ ⟶ D
	Line 7A to line 4	None	—
② to ⑤	Line 1 to line 2	⑪	O ⟶ ② ⟶ ⑪ ⟶ ⑤ ⟶ D
	Line 1 to line 7B	None	—
	Line 1 to line 7C	⑥	O ⟶ ② ⟶ ⑥ ⟶ ⑤ ⟶ D
	Line 7A to line 2	①	O ⟶ ② ⟶ ① ⟶ ⑤ ⟶ D
	Line 7A to line 7B	③	O ⟶ ② ⟶ ③ ⟶ ⑤ ⟶ D
	Line 7A to line 7C	None	—

Table 11: Travel times, transfer times, number, and broad cost of transfers for the seven lines.

No	Path	Travel time (s)	Transfer time (s)	Number of transfers (time)	Broad cost
1	O ⟶ ① ⟶ ⑤ ⟶ D	2498	158	2	3228.7
2	O ⟶ ① ⟶ ⑨ ⟶ ④ ⟶ D	2574	324	3	5762.4
3	O ⟶ ① ⟶ ⑧ ⟶ ④ ⟶ D	2490	136	2	3119.0
4	O ⟶ ② ⟶ ⑩ ⟶ ④ ⟶ D	2519	355	3	6012.5
5	O ⟶ ② ⟶ ③ ⟶ ④ ⟶ D	1892	0	0	1892.0
6	O ⟶ ② ⟶ ⑪ ⟶ ⑤ ⟶ D	2841	321	3	5999.9
7	O ⟶ ② ⟶ ⑥ ⟶ ⑤ ⟶ D	2575	115	2	3106.9

Figure 9 shows the Chengdu subway network after our numbering process, where the line numbers follow the operating line numbers except for Line 7, and the black circles represent interchange stations. The numbering is discontinuous because this example focuses on the lines within the loop of Line 7 while omitting Line 10 and the branch of Line 1 from Sihe to Wugensong, which have no line crossings. In Figure 9, Line 7 is a loop and contains several valid paths such as direct and detour in the path from point O to point D. One of the bypass paths violates the above valid path assumption (4), so we break the line containing arcs or loops into several branches for path passenger distribution at appropriate places. In Figure 9, the interrupted stations are numbered 1, 3, 5, and 7, corresponding to the operating stations of Yipin World, Yima Bridge, Chengdu East Passenger Station, and Taiping Park, forming lines 7A, 7B, 7C, and 7D, respectively.

The following is the search process for the set of valid paths based on our proposed "two-way search algorithm."

Step 1: Determine the adjacent interchanges or terminal stations at points O and D, respectively. Based on the subordinate relationship between the line and the station, i.e., $L_i = \{v_1, v_2, \cdots, v_k\}$ (L_i for line, v_k for station), the adjacent interchange stations at the original point O can be determined as ① and ②, and the adjacent interchange stations at the ending point D as ④ and ⑤.

Step 2: Determine the lines where each adjacent interchange is located separately. Interchange ① is subordinate to Line 2, Line 7A, and Line 7D; Interchange ② is subordinate to Line 1 and Line 7A; Interchange ④ is subordinate to Line 4 and Line 7B; Interchange ⑤ is subordinate to Line 2, Line 7B and Line 7C.

Step 3: Crossover determines whether each vector interchange station is on the same line. Stations ① and ④, stations ② and ④, and stations ② and ⑤ are not on the same line, so a further search for interchange stations is needed. Both stations ① and ⑤ are located online 2, so the route O ⟶ ① ⟶ ⑤ ⟶ D can be determined.

Step 4: Search for a valid path containing three interchanges. The search process is described in Table 10:

Step 5: According to steps three and four, remove the paths containing duplicate segments to obtain the final set of valid paths and complete the search. The final set of valid paths contains the following 7 entries, as shown in. Paths 3, 5, and 7 contain virtual interchange arcs (7D, 7B, etc.), thus reducing the number of interchanges compared to the representation of paths. Substituting the above metrics into the route selection model, the broad cost values are obtained in Table 11:

Since the effective paths searched by the above algorithm are still relatively large and the paths considered by urban rail passengers are often only 1~3, the travel time, transfer time, and number of transfers for each path are calculated by further reducing the stretch factor H of the path. According to the stretch factor H of the path, when the broad cost of a path is greater than $(1 + H)$ times the minimum broad cost, the path will not be considered by the traveler and should be removed from the set of valid paths. When H takes the calibration result of 0.25, the path with the smallest broad cost is valid path5, and all other paths are eliminated. After substituting the path selection model, the passenger flow matching probability of path O ⟶ ② ⟶ ③ ⟶ ④ ⟶ D is 100%, i.e., according to our algorithm, all passengers travelling between Cha Dian Zi Station and Ying Hui Road Station will choose Line 7 directly as the only path.

Further analysis reveals that path 5 contains three virtual transfer arcs (line 7A and switching line 7B) with the shortest travel time, zero transfer time, and zero number of transfers, thus meeting the actual path selection willingness of travelers. In the other OD selection cases, there are multiple effective paths with closer broad costs, and multiple path selection results with less than 1 allocation ratio can be obtained by our proposed method. In summary, our passenger path assignment algorithm largely proves to be accurate and effective.

5. Conclusions

Our research aims to predict urban rail traffic, specifically in terms of the destination stations and travel routes that commuters will choose. To achieve this, we focused on commuter traffic as our research object, as it has a high proportion and strong travel regularity. We utilized passenger entry information from rail transit stations with a high proportion of commuter traffic, and our contributions are outlined below. First, we divided passenger flow into two categories based on the formation of travel habits and performed OD prediction using a combination of data mining and logit modeling. As passenger flow can be unstable, we split the flow into passengers who have formed travel habits and those who have not yet formed these habits. For the first group, we utilized a mining algorithm based on historical travel habits to predict their travel destinations using historical AFC data. For those who have not formed travel habits, we mainly used a modified ML model to predict the most likely outbound station a passenger will choose when entering a station, considering spatiotemporal influences such as travel time, regional attractiveness, and OD size. Second, determining a passenger's choice path between two points based on OD is a key step that requires designing efficient algorithms to find complete and effective paths. To do this, we assumed that the number of interchanges would not exceed that when passengers chose a route and that there was an effective route. Using a "two-way search algorithm," we searched adjacent interchange stations and line interchange stations from the origin and destination of the OD pair at the same time, making full use of interchange stations to implement network topology modeling. This approach allowed us to quickly search for a complete and effective route, which we verified through experiments. Last, our algorithm exhibits good generality and can be applied to rail transportation networks in different cities. The forecasting model that we developed is a service for urban rail operators and passengers who use urban rail to travel. Our model aims to provide a holistic forecast of commuter flow in terms of travel stations and tracking and analyzing the travel destinations of each passenger. In addition, it provides detour information for traffic participants to avoid congested stations and supports decision-makers on current and next-period passenger flow conditions to respond to unexpected situations. While our research has made important contributions, some problems can still not be solved due to limited capacity. For example, in the final example analysis, we used a dummy variable to mark the influence factor of regional attractiveness. In future studies, a distribution function could be introduced to quantify regional attractiveness.

Data Availability

The data used to support the findings of this study are included in the article. Should further data or information be required, these are available from the corresponding author upon request.

Conflicts of Interest

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

Acknowledgments

This work was supported by the National Science Foundation of Liaoning Province, China, under Grant No. 2023-MS-273.

References

[1] Z. Guo, X. Zhao, Y. Chen, W. Wu, and J. Yang, "Short-term passenger flow forecast of urban rail transit based on GPR and KRR," *IET Intelligent Transport Systems*, vol. 13, no. 9, pp. 1374–1382, 2019.

[2] L. Tang, Y. Zhao, J. Cabrera, J. Ma, and K. Tsui, "Forecasting short-term passenger flow: an empirical study on shenzhen metro," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 10, pp. 3613–3622, 2019.

[3] Y. Yang, J. Liu, P. Shang, X. Xu, and X. Chen, "Dynamic origin–destination matrix estimation based on urban rail transit AFC data: deep optimization framework with forward passing and backpropagation techniques," *Journal of Advanced Transportation*, vol. 2020, Article ID 8846715, 16 pages, 2020.

[4] Y. Cao, X. Hou, and N. Chen, "Short-term forecast of OD passenger flow based on ensemble empirical mode decomposition," *Sustainability*, vol. 14, no. 14, p. 8562, 2022.

[5] X. Yao, P. Zhao, and D. Yu, "Real-time Origin–Destination matrices estimation for urban rail transit network based on

structural state-space model," *Journal of Central South University*, vol. 22, no. 11, pp. 4498–4506, 2015.

[6] J. Wang, K. Liu, T. Yamamoto, D. Wang, and G. Lu, "Built environment as a precondition for demand-responsive transit (DRT) system survival: evidence from an empirical study," *Travel Behaviour and Society*, vol. 30, pp. 271–280, 2023.

[7] J. Lu, G. Ren, and L. Xu, "Analysis of subway station distribution capacity based on automatic Fare collection data of nanjing metro," *Journal of Transportation Engineering Part A-Systems*, vol. 146, no. 2, 2020.

[8] B. Du, Y. Cui, Y. Fu, R. Zhong, and H. Xiong, "SmartTransfer: modeling the spatiotemporal dynamics of passenger transfers for crowdedness-aware route recommendations," *ACM Transactions on Intelligent Systems and Technology*, vol. 9, no. 6, pp. 1–26, 2018.

[9] W. Chen, Z. Li, C. Liu, and Y. Ai, "A deep learning model with conv-LSTM networks for subway passenger congestion delay prediction," *Journal of Advanced Transportation*, vol. 2021, Article ID 6645214, 10 pages, 2021.

[10] L. Liu, J. Chen, H. Wu, J. Zhen, G. Li, and L. Lin, "Physical-virtual collaboration modeling for intra- and inter-station metro ridership prediction," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 4, pp. 3377–3391, 2022.

[11] J. Yang, X. Han, T. Ye et al., "Spatiotemporal virtual graph convolution network for key origin–destination flow prediction in metro system," *Mathematical Problems in Engineering*, vol. 2022, Article ID 5622913, 11 pages, 2022.

[12] J. Ye, J. Zhao, F. Zheng, and C. Xu, "Completion and augmentation-based spatiotemporal deep learning approach for short-term metro origin-destination matrix prediction under limited observable data," *Neural Computing and Applications*, vol. 35, no. 4, pp. 3325–3341, 2022.

[13] F. Yang, C. Shuai, Q. Qian et al., "Predictability of short-term passengers' origin and destination demands in urban rail transit," *Transportation*, vol. 50, no. 6, pp. 2375–2401, 2022.

[14] Y. Zhang and E. Yao, "Splitting travel time based on AFC data: estimating walking, waiting, transfer, and in-vehicle travel times in metro system," *Discrete Dynamics in Nature and Society*, vol. 2015, Article ID 539756, 11 pages, 2015.

[15] J. Zeng and J. Tang, "Combining knowledge graph into metro passenger flow prediction: a split-attention relational graph convolutional network," *Expert Systems with Applications*, vol. 213, Article ID 118790, 2023.

[16] M. Aqib, R. Mehmood, A. Alzahrani, I. Katib, A. Albeshri, and S. Altowaijri, "Rapid transit systems: smarter urban planning using big data, in-memory computing, deep learning, and GPUs," *Sustainability*, vol. 11, no. 10, pp. 2736–2833, 2019.

[17] X. Yang, Q. Xue, M. Ding, J. Wu, and Z. Gao, "Short-term prediction of passenger volume for urban rail systems: a deep learning approach based on smart-card data," *International Journal of Production Economics*, vol. 231, Article ID 107920, 2021.

[18] P. Noursalehi, H. Koutsopoulos, and J. Zhao, "Dynamic origin–destination prediction in urban rail systems: a multi-resolution spatio-temporal deep learning approach," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 6, pp. 5106–5115, 2022.

[19] Z. Cheng, M. Trepanier, and L. Sun, "Real-time forecasting of metro origin–destination matrices with high-order weighted dynamic mode decomposition," *Transportation Science*, vol. 56, no. 4, pp. 904–918, 2022.

[20] Z. Cai, T. Li, X. Su, L. Guo, and Z. Ding, "Research on analysis method of characteristics generation of urban rail transit," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 9, pp. 3608–3620, 2020.

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

[22] E. Yao, J. Hong, L. Pan, B. Li, Y. Yang, and D. Guo, "Forecasting passenger flow distribution on holidays for urban rail transit based on destination choice behavior analysis," *Journal of Advanced Transportation*, vol. 2021, Article ID 9922660, 13 pages, 2021.

[23] Y. Zhu, H. Koutsopoulos, and N. Wilson, "Passenger itinerary inference model for congested urban rail networks," *Transportation Research Part C: Emerging Technologies*, vol. 123, Article ID 102896, 2021.

[24] J. Wu, Y. Qu, H. Sun, H. Yin, X. Yan, and J. Zhao, "Data-driven model for passenger route choice in urban metro network," *Physica A: Statistical Mechanics and its Applications*, vol. 524, pp. 787–798, 2019.

[25] W. Jiang, Z. Ma, and H. Koutsopoulos, "Deep learning for short-term Origin–Destination passenger flow prediction under partial observability in urban railway systems," *Neural Computing and Applications*, vol. 34, no. 6, pp. 4813–4830, 2022.

[26] J. Zhang, H. Che, F. Chen, W. Ma, and Z. He, "Short-term Origin–Destination demand prediction in urban rail transit systems: a channel-wise attentive split-convolutional neural network method," *Transportation Research Part C: Emerging Technologies*, vol. 124, Article ID 102928, 2021.

[27] L. Shen, Z. Shao, Y. Yu, and X. Chen, "Hybrid approach combining modified gravity model and deep learning for short-term forecasting of metro transit passenger flows," *Transportation Research Record*, vol. 2675, no. 1, pp. 25–38, 2021.

[28] C. Li, J. Huang, B. Wang, Y. Zhou, Y. Bai, and Y. Chen, "Spatial-temporal correlation prediction modeling of origin–destination passenger flow under urban rail transit emergency conditions," *IEEE Access*, vol. 7, pp. 162353–162365, 2019.

[29] G. Zhu, J. Ding, Y. Wei, Y. Yi, S. Xu, and E. Wu, "Two-stage OD flow prediction for emergency in urban rail transit," *IEEE Transactions on Intelligent Transportation Systems*, vol. 2023, Article ID 3235413, 9 pages, 2023.

[30] X. Wang, C. Zhu, and J. Jiang, "A deep learning and ensemble learning based architecture for metro passenger flow forecast," *IET Intelligent Transport Systems*, vol. 17, no. 3, pp. 487–502, 2022.

[31] C. Yu, H. Li, X. Xu et al., "Data-Driven approach for passenger mobility pattern recognition using spatiotemporal embedding," *Journal of Advanced Transportation*, vol. 2021, Article ID 5574093, 21 pages, 2021.

[32] O. Kosheleva, V. Kreinovich, and S. Sriboonchitta, "Econometric models of probabilistic choice: beyond McFadden's formulas," in *Robustness in Econometrics*, V. Kreinovich, S. Sriboonchitta, and V.-N. Huynh, Eds., vol. 692, pp. 79–87, Springer International Publishing, Cham, Switzerland, 2017.

[33] R. Dial, "Transit pathfinder algorithm," *Highway Research Record*, vol. 205, pp. 67–85, 1967.

[34] M. G. H. Bell, "Alternatives to Dial's logit assignment algorithm," *Transportation Research Part B: Methodological*, vol. 29, no. 4, pp. 287–295, 1995.

[35] J. Arriagada, M. A. Munizaga, C. A. Guevara, and C. Prato, "Unveiling route choice strategy heterogeneity from smart card data in a large-scale public transport network,"

Transportation Research Part C: Emerging Technologies, vol. 134, Article ID 103467, 2022.

[36] Y. Liu, T. Feng, Z. Shi, and M. He, "Understanding the route choice behaviour of metro-bikeshare users," *Transportation Research Part A: Policy and Practice*, vol. 166, pp. 460–475, 2022.

[37] B. J. Tomhave and A. Khani, "Refined choice set generation and the investigation of multi-criteria transit route choice behavior," *Transportation Research Part A: Policy and Practice*, vol. 155, pp. 484–500, 2022.

[38] K. Ma, J. Wen, and Q. Wang, "Rail transit travel time distribution and prediction based on automatic Fare collection data," *Proceedings of the Beijing Jiaotong University*, vol. 37, pp. 120–123, 2016.

[39] D. Li, T. Miwa, C. Xu, and Z. Li, "Non-linear fixed and multilevel random effects of origin–destination specific attributes on route choice behaviour," *IET Intelligent Transport Systems*, vol. 13, no. 4, pp. 654–660, 2019.

[40] W. Wu, D. Li, and C. Li, "Estimated OD matrix based on multipath choice model by using urban traffic flow," in *Proceedings of the 19th COTA International Conference of Transportation Professionals*, pp. 5492–5502, Nanjing, China, June 2019.