Journal of Advanced Transportation

Big Data in Public Transport Operation

Lead Guest Editor: Liu Jun
Guest Editors: Haiying Li, Linchao Li, Lingqiao Qin, and Xinyue Xu

Big Data in Public Transport Operation

Journal of Advanced Transportation

Big Data in Public Transport Operation

Lead Guest Editor: Liu Jun
Guest Editors: Haiying Li, Linchao Li, Lingqiao Qin, and Xinyue Xu

Associate Editors

Academic Editors

Contents

Hindawi
Journal of Advanced Transportation
Volume 2022, Article ID 6657486, 14 pages
https://doi.org/10.1155/2022/6657486

Research Article

Exploring for Route Preferences of Subway Passengers Using Smart Card and Train Log Data

Eun Hak Lee,[1] **Kyoungtae Kim**,[2] **Seung-Young Kho**,[1,3] **Dong-Kyu Kim**,[1,3] **and Shin-Hyung Cho**[4]

[1] *Institute of Construction and Environmental Engineering, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul 08826, Republic of Korea*
[2] *Future Transport Policy Research Division, Korea Railroad Research Institute, 176, Cheoldobangmulgwan-ro, Uiwang-si, Gyeonggi-do 16106, Republic of Korea*
[3] *Department of Civil and Environmental Engineering, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul 08826, Republic of Korea*
[4] *School of Civil and Environmental Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA*

Correspondence should be addressed to Shin-Hyung Cho; scho370@gatech.edu

Received 18 December 2020; Revised 12 October 2021; Accepted 10 November 2021; Published 3 January 2022

Academic Editor: Hai Ying Li

As the mode share of the subway in Seoul has increased, the estimation of passenger travel routes has become a crucial issue to identify the congestion sections in the subway network. This paper aims to estimate the travel train of subway passengers in Seoul. The alternative routes are generated based on the train log data. The travel route is then estimated by the empirical cumulative distribution functions (ECDFs) of access time, egress time, and transfer time. The train choice probability is estimated for alternative train combinations and the train combination with the highest probability is assigned to the subway passenger. The estimated result is validated using the transfer gate data which are recorded on private subway lines. The result showed that the accuracy of the estimated travel train is shown to be 95.6%. The choice ratios for no-transfer, one-transfer, two-transfer, three-transfer, and four-transfer trips are estimated to be 53.9%, 37.7%, 6.5%, 1.5%, and 0.4%, respectively. Regarding the practical application, the passenger kilometers by lines are estimated with the travel route estimation of the whole network. As results of the passenger kilometer calculation, the passenger kilometer of the proposed algorithm is estimated to be 88,314 million passenger kilometer. The proposed algorithm estimates the passenger kilometer about 13% higher than the shortest path algorithm. This result implies that the passengers do not always prefer the shortest path and detour about 13% for their convenience.

1. Introduction

In 2004, the municipal government of Seoul introduced the automatic fare collection (AFC) system. The AFC system makes it possible to analyze the travel behavior of transit passengers. With smart card data obtained from the AFC system, it has much attention to estimate the travel route of passengers on subway networks [1]. Seoul's transit fare system charges passengers based on their travel distance, so it is essential to ascertain the passenger's travel routes [2]. Smart card data of the AFC system provide travel route information of bus trips and transfer trips between the bus and subway networks [3, 4]. The travel routes of the subway passengers, however, are still hard to identify since the smart card data do not provide route information of subway passengers [5]. The card reader of the subway AFC system is installed at the gates of the station, which is outside of the platform. Since the information is only recorded at the station gates that a passenger departs or arrives, thus there is no way to know which route a passenger has traveled. The crucial problem of estimating the travel routes of subway passengers is that there is no information about transfer trips between public subway lines [6]. Only privately owned lines have installed the transfer gates, which are located on the transfer aisle. Travel route information of trips made through the private lines can be identified with the transfer gate data. For the transfer trips of public lines, the travel route information is not provided since there is no transfer gate at the transfer station.

The travel routes of urban railways have traditionally been estimated based on utility maximization or regret minimization models [7, 8]. However, these models could not be valid for several reasons. The train arrival time is not always consistent with the train schedule in a complex urban railway system. Also, passengers might not choose the estimated travel route depending on their tap-in time and train arrival time. Passengers could choose unexpected travel routes with instantaneous decisions. Thus, the traditional models were not always correct in these specific situations, and the advanced method is required to estimate the travel route [9].

Recently, many studies have explored route preference using smart card data [10–14]. For example, Sun et al. [15] estimated the passenger's location with smart card data of the Singapore MRT system. The spatiotemporal density of passengers was estimated, and the trains' trajectories were identified from the move of estimated density. These results were derived from the railway network in which consecutive trains followed the same route without transfers. Similarly, Kusakabe et al. [16] explored the passenger's train choice behavior with smart card data. The route with the longest in-vehicle time was selected as the traveled route rather than the earliest departing or arriving routes. Lee et al. [17] also estimated the express train choice behavior using smart card data. The Gaussian mixture model was used to decompose the travel time distribution into two distributions, i.e., express train and local train. Each passenger was assigned to an express or local train according to a density probability.

Many previous studies have sought to accurately explore passenger's train preferences using smart card data and train log data, i.e., train logs or train schedules [18, 19]. For example, Sun and Xu [20] estimated the egress time, access time, transfer time, and in-vehicle time with the smart card data, train schedules, and complementary manual surveys. With these estimated attributes, the travel time distribution of each route was established, and the passenger preference was explored. Zhou and Xu [13] also estimated the traveled route to assign passenger flow. With the train schedule data, feasible routes were generated, and each passenger was assigned to the route, which had a minimum surplus time. Similarly, Zhu et al. [21] estimated the train choice behavior with real timetables and smart card data. The choice set was generated by the deletion algorithm, and the route choice probability was estimated by Manski's paradigm. Sun and Schonfeld [22] proposed a route choice model using smart card data. The choice set was generated based on the train schedule connection network. The access time, egress time, and transfer time were considered to assign passengers to the generated route. Similarly, Hong et al. [23] also proposed a train choice model with smart card data and train log data. The passengers who have a unique route were defined as reference passengers, and the traveled routes of passengers who have multi-route were estimated by matching the reference passengers.

Although these previous studies attempted to estimate the travel route, some improvements still remained. First, the accuracy of the route estimation needed to be improved using passenger's experienced travel time attributes, i.e., access time, egress time, transfer time, and in-vehicle time. The distribution forms of the travel time attributes are all different by stations and origin-destination (O-D) pairs. Thus, travel time attributes are required to estimate without the distribution assumption. Second, there was a limitation on validating the model performance since passenger's travel route information, such as transfer information, was not recorded on smart card data. Previous studies have proposed many methods to estimate travel routes. However, there is a limit to identifying the accuracy of the method due to the absence of revealed preference data of travel routes. To shed light on these issues, this study proposed a methodology that estimates passenger's travel route (train) using smart card data and train log data. The contributions of this study were presented as follows: (1) the empirical distribution without distribution assumption was developed to estimate the probability of each travel time attribute; (2) model performance was validated with revealed route information (transfer gate) data; and (3) the practical application, such as efficiency evaluation of each subway line, was performed using estimated results of the whole subway passengers in Seoul.

This study estimated the travel route of individual subway passengers using the smart card data and train log data. The alternative routes were generated based on the train log data. The travel route was then estimated by the empirical cumulative distribution functions (ECDFs) of access time, egress time, and transfer time. With the ECDFs of the time attributes, the train choice probability was estimated for alternative train combinations. Among the alternative train combinations, the train combination with the highest probability was assigned to the subway passenger. The smart card data of the private lines were employed to validate the results of the travel train estimation since it had the exact information about the travel route transaction. The proposed algorithm was then applied to estimate the travel train of all subway passengers on the entire subway network in Seoul.

2. Data Description

2.1. Description of the Network (Seoul Metropolitan Area). The subway network in Seoul consists of 11 lines numbering from 1 to 9, Bundang Line, and Shinbundang Line. The subway network has 327 stations, including 127 transfer stations to serve Seoul and its surroundings. Among 11 lines, Line 9 and the Shinbundang Line are owned by private companies. The total number of trips of the subway network in Seoul is 6,313,176 trips per day. The headway of the subway trains is about 6 minutes on average. The minimum and maximum headways are about 2 and 26 minutes, respectively. There is no way to identify the travel route with the public lines. However, private lines have transfer gates at all transfer stations to collect fares. With the data from the transfer gate, it is possible to validate the results of the travel route estimation. Line 9 consists of 30 stations with nine transfer stations, and the Shinbundang Line consists of 12 stations with five transfer stations. The number of trips of Line 9 and Shinbundang Line is 472,436 trips per day. Since the percentage of private trips accounts for about 7.4% of all trips, it is possible to validate the estimation result.

The travel route estimation for the trips traveled private lines was conducted to validate the performance of the proposed algorithm. The process of estimating train choices for the individual passenger was explained with an illustration network that has two alternative routes for the same O-D pair. The travel route for the subway network in Seoul was also estimated to ascertain the practical applicability of the algorithm. The subway network in Seoul is shown in Figure 1.

2.2. Descriptions of the Smart Card Data and Train Log Data. The smart card data store about 20 million trip information per day, including about 7 million subway trips and 12 million bus trips. The smart card data can be obtained from the Korea Transportation Safety Authority (KTSA) and contain 38 data information for each trip. To estimate the train choice, we used smart card data of October 31, 2017. Among the 38 data information, we used 10; card ID, transaction ID, line ID, boarding station ID, alighting station ID, boarding time, alighting time, total travel time, transfer station ID, and transfer time. The data information related to the transfer is provided only from the trips on the two private lines. Thus, it is possible to identify the travel route of passengers who traveled on private lines. The data information of the smart card data are shown in Table 1.

The train log data contain about 175,000 logs of real-time train operation data per day. The train log data can be obtained from the Open Data Portal (data.seoul.go.kr), and it includes the arrival time information of the train at each station. The reliability of the train log data is ensured because it is the actual arrival time of the train. By integrating train log data with the smart card data, it is possible to estimate the passenger travel route. The train log data used in this study are also from October 31, 2017. It contains eight data information, of which seven data information were used: line ID, arrival time, the direction of train, train ID, train type, boarding station ID, and alighting station ID. The data information of train log data is shown in Table 2.

3. Methodology

The proposed train choice algorithm has two main methodologies, i.e., choice set generation algorithm and empirical cumulative distribution functions (ECDFs). The choice set generation algorithm is used to generate the available train combinations for each passenger. The ECDFs methodology is used to estimate the passenger's choice probability for each alternative. The proposed train choice algorithm consists of seven steps using a choice set generation algorithm and ECDFs. The visualized concept of the train choice algorithm and definition of notations are shown in Figure 2 and Table 3, respectively. For a better understanding of the proposed train choice algorithm, the remainder of the methodology section is organized as follows: the concept of choice set generation algorithm and the concept of ECDFs is described in order. Then, the seven steps of the proposed train choice model are explained step by step.

3.1. Choice Set Generation. In this part, we proposed an algorithm to generate alternative train combinations for an individual passenger using the tap-in time and tap-out time of smart card data, and train arrival time of train log data. The alternative train combination connects the passenger's origin and destination stations during his/her travel time. With the proposed algorithm, it is possible to generate all train choice alternatives for each subway passenger.

The choice set generation is performed for each passenger. Thus, alternative train combinations could be different for the passengers even with the same origin to destination (O-D). The proposed algorithm considered all alternative routes using alternative train combinations during the passenger's travel time. choice combinations during the passenger's travel time. The mathematical expression of the algorithm of generating the alternative train combination is shown in equations (1) to (4). Equation (3) is to find all available trains which depart the origin and arrive at the destination stations between the tap-in and tap-out times of an individual passenger. If there is a transfer station, the train choice combination is generated by connecting transferable trains and the available trains. Equation (4) shows the mathematical expression of the alternative train combination set of the trip i.

$$N = \{1, 2, 3, \ldots, 363\}, \tag{1}$$

$$\mathbf{p}_i = \left(t_i^{\text{in}}, t_i^{\text{out}}, o_i, d_i\right), \quad o \in N, d \in N, \tag{2}$$

$$\mathbf{r} = \left(\delta, tr_{\text{in}}^{1}, tr_{\text{out}}^{1}, tr_{\text{in}}^{2}, tr_{\text{out}}^{2}, \ldots, tr_{\text{in}}^{k}, tr_{\text{out}}^{k}, \alpha, o, d\right), \quad o \in N, d \in N, \tag{3}$$

$$R_{OD}(\mathbf{p}_i) = \left\{\mathbf{r} | \delta \geq t_i^{\text{in}}, \alpha \leq t_i^{\text{out}}, o = o_i, d = d_i\right\}. \tag{4}$$

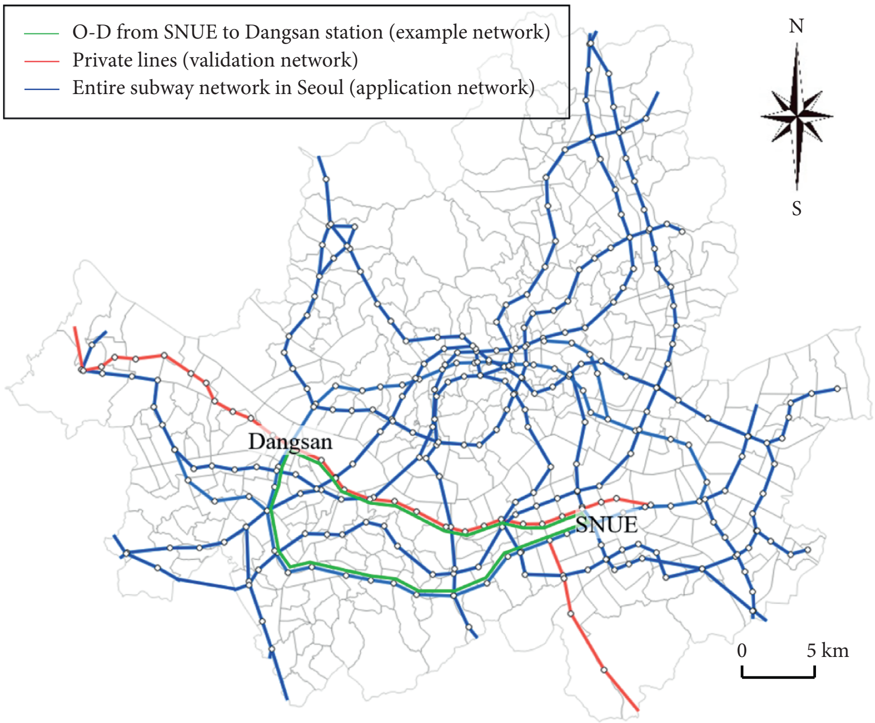

O-D from SNUE to Dangsan station (example network)
Private lines (validation network)
Entire subway network in Seoul (application network)
N
S
Dangsan
SNUE
0
5 km

Figure 1: Subway network in Seoul.

3.2. Empirical Cumulative Distribution Function. The ECDF is a nonparametric estimator of the typical CDF of a random variable. ECDF has an advantage in estimating probabilities because assumptions are relatively free. For example, distributions of the travel time attributes are difficult to define in the specific form since the distribution of each station and O-D pairs is all different. If there are plenty of samples, the ECDF can improve the accuracy of the model. In other words, the ECDF approximates the true CDF with the large samples. It estimates a probability of $1/j$ to each sample, orders the samples from smallest to largest in value, and calculates the sum of the estimated probabilities up to and including each sample value. The result is a step function that increases by $1/j$ at each sample value. The ECDF is usually denoted by f_j or $P_j(X \le x)$, and mathematical expression is defined as follows:

$$f_j(x) = P_j(X \le x) = j^{-1}\sum_{i=1}^{j} I(x_i \le x). \tag{5}$$

$I(x_p \le x)$ is the indicator function and has two values. If the event inside the brackets occurs, the value is 1, and if not, the value is 0.

$$I(x_p \le x) = \begin{cases} 1, & \text{when } x_p \le x, \\ 0, & \text{when } x_p > x. \end{cases} \tag{6}$$

3.3. Train Choice Algorithm. To estimate the passengers' travel train combinations, we developed a train choice algorithm using smart card data and train log data. The proposed algorithm consists of seven steps. Step 1 is to extract information about passengers who have a clear train combination to travel. In this case, the passenger has only one train available to travel from the origin station to the destination station between tap-in time and tap-out time. In Step 2, the time attributes, i.e., access time, egress time, and transfer time, are calculated by the extracted passenger's tap-in time and tap-out time and train arrival time and departure time. In Step 3, the ECDFs of access time, egress time, and transfer time for each station are developed using the calculated time attributes. Step 4 is for generating alternative train choices for a passenger who has more than two alternative trains on his/her route. In Step 5, the choice probability is estimated for each alternative train. The train choice probability is calculated by multiplying the probability of time attribute, i.e., access time, egress time, and transfer time for all of the alternative trains. The probability of each travel time attribute converges to 1 as it approaches the mode value. In step 6, the train combination with the highest choice probability is assigned to a passenger. Step 7 is the iteration step for estimating the next passenger's travel train combination. The mathematical expression of the travel train estimation algorithm is shown in equations (7) to (19).

Step 1. Select the set of passengers who have only one alternative train combination during his/her travel time.

The passenger group with one train available is selected by comparing the tap-in time and tap-out time of smart card data to the train arrival time at the origin station of the train log data. Specifically, all available train combinations during the tap-in time and tap-out time of each passenger are

Table 1: Description of the smart card data.

No.	Data information
1	Card ID*
2	Transaction ID*
3	Mode code
4	Line ID*
5	Name of the transit line
6	Vehicle ID
7	Vehicle number
8	Boarding station ID*
9	Alighting station ID*
10	Name of boarding station
11	Name of alighting station
12	Boarding (tap-in) time*
13	Alighting (tap-out) time*
14	Number of transfer
15	Total travel distance
16	Total travel time*
17	Boarding fare
18	Alighting fare
19	The number of users
20	Boarding violation penalty
21	Alighting violation penalty
22	General user code
23	Student user code
24	Child user code
25	Other user code
26	User division
27	User group
28	Company code
29	Company name
30	Time code
31	Starting run time
32	Ending run time
33	Boarding date
34	Alighting date
35	Year
36	Zone code
37	Transfer station ID
38	Transfer time

*Used in this study.

checked, and a passenger who has only one available train is selected in this step.

$$U = \{\mathbf{p}_i | \mathrm{n}(R_{OD}(\mathbf{p}_i)) = 1, \quad \mathbf{p}_i \in P = \{\mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_3, \dots, \mathbf{p}_i \dots, \mathbf{p}_{6,313,176}\}\}. \tag{7}$$

Step 2. Calculate the travel time attributes of the set of passengers who have only one alternative train combination.

The access time, the egress time, and the transfer time of individual passengers are estimated using the tap-in time and tap-out time from the smart card data and train arrival time at the origin, transfer, and destination stations.

$$a_i = \delta - t_i^{\mathrm{in}}, \tag{8}$$

$$e_i = t_i^{\mathrm{out}} - \alpha, \tag{9}$$

Table 2: Description of the train log data.

No.	Data information
1	Name of affiliate
2	Line ID
3	Arrival time
4	The direction of the train
5	Train ID
6	Train type
7	Boarding station ID
8	Alighting station ID

$$tr_i = tr_{\mathrm{out}}^k - tr_{\mathrm{in}}^k. \tag{10}$$

Subject to

$$r = \left(\delta, tr_{\mathrm{in}}^1, tr_{\mathrm{out}}^1, tr_{\mathrm{in}}^2, tr_{\mathrm{out}}^2, \dots, tr_{\mathrm{in}}^k, tr_{\mathrm{out}}^k, \alpha, o, d\right) \in R_{OD}(\mathbf{p}_i). \tag{11}$$

$$\mathbf{p}_i = \left(t_i^{\mathrm{in}}, t_i^{\mathrm{out}}, o_i, d_i\right) \in U. \tag{12}$$

Step 3. Develop the empirical cumulative distribution function (ECDF) of time attributes.

ECDFs are set up using the access time, the egress time, and the transfer time of individual passengers who have only one train available.

$$F_a^o(a_u^o) = f_j(a_u^o), \quad \text{for } u \text{ s.t. } \mathbf{p}_u \in U, \tag{13}$$

$$F_e^d\left(e_u^d\right) = f_j\left(e_u^d\right), \quad \text{for } u \text{ s.t. } \mathbf{p}_u \in U, \tag{14}$$

$$F_{tr}^k\left(tr_u^k\right) = f_j\left(tr_u^k\right), \quad \text{for } u \text{ s.t. } \mathbf{p}_u \in U. \tag{15}$$

Step 4. Generate alternative train combinations for a passenger who has multiple alternatives.

The set of passengers could be generated when they have multiple trains available at origin, transfer, and destination stations between their tap-in time and tap-out time.

$$M = \{\mathbf{p}_i | \mathrm{n}(R_{OD}(\mathbf{p}_i)) > 1, \quad \mathbf{p}_i \in P = \{\mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_3, \dots, \mathbf{p}_i \dots, \mathbf{p}_{6,313,176}\}\}. \tag{16}$$

Step 5. Calculate the choice probability of each alternative train.

The choice probability of each alternative train was estimated by multiplying three probabilities of access time, transfer time, and egress time. The probability of the mode value was assumed to be 100% since the travel time attributes formed the skewed distribution. As the travel time attributes become closer to the mode value, there will get a higher chance to board the train. Therefore, the probability was defined based on the distance from the mode value as the probability of the corresponding time attributes.

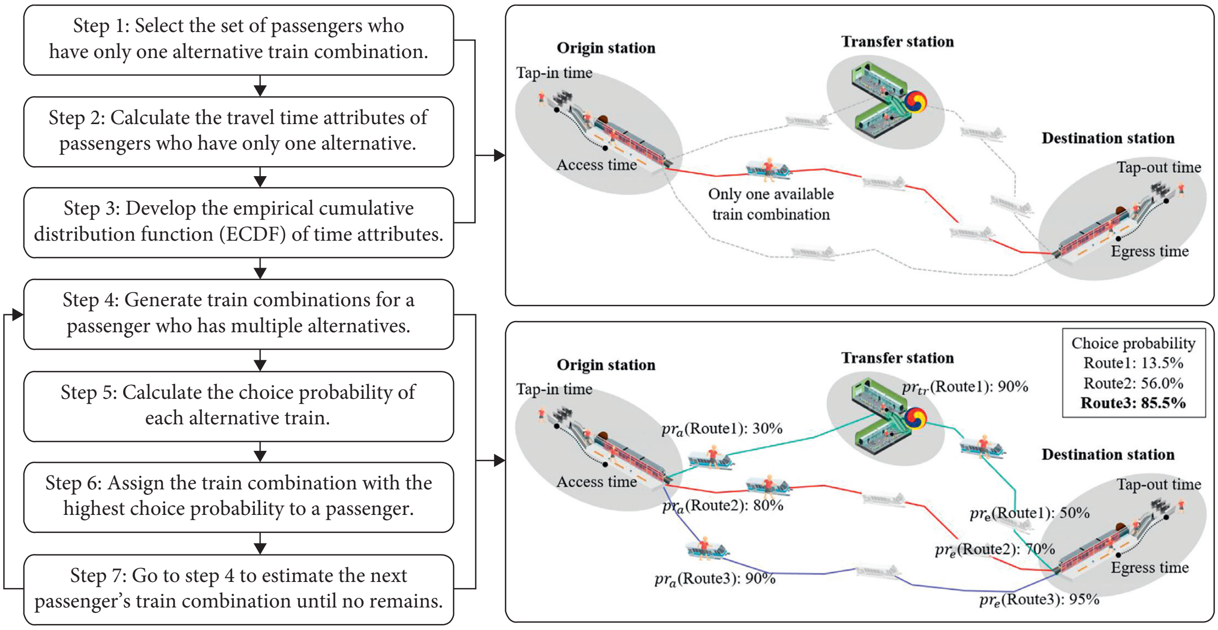

Step 1: Select the set of passengers who have only one alternative train combination.
Step 2: Calculate the travel time attributes of passengers who have only one alternative.
Step 3: Develop the empirical cumulative distribution function (ECDF) of time attributes.
Step 4: Generate train combinations for a passenger who has multiple alternatives.
Step 5: Calculate the choice probability of each alternative train.
Step 6: Assign the train combination with the highest choice probability to a passenger.
Step 7: Go to step 4 to estimate the next passenger's train combination until no remains.
Origin station
Tap-in time
Access time
Transfer station
Only one available train combination
Destination station
Tap-out time
Egress time
Origin station
Tap-in time
Access time
Transfer station
pr_{tr}(Route1): 90%
Choice probability
Route1: 13.5%
Route2: 56.0%
Route3: 85.5%
pr_a(Route1): 30%
pr_a(Route2): 80%
pr_a(Route3): 90%
Destination station
Tap-out time
pr_e(Route1): 50%
pr_e(Route2): 70%
Egress time
pr_e(Route3): 95%

Figure 2: Visualized concept of train choice algorithm.

$$pr = pr_a * pr_e * pr_{tr}, \tag{17}$$

$$pr_a = 1 - \left| F_a^o\left(a_m^o\right) - F_a^o\left(ma_u^o\right) \right|, \tag{18}$$

$$pr_e = 1 - \left| F_e^d\left(e_m^d\right) - F_e^d\left(me_u^d\right) \right|, \tag{19}$$

$$pr_{tr} = 1 - \left| F_{tr}^k\left(tr_m^k\right) - F_{tr}^k\left(mtr_u^k\right) \right|, \quad \text{for } u \text{ s.t. } \mathbf{p_u} \in U \text{ for } m \text{ s.t. } \mathbf{p}_m \in M. \tag{20}$$

Step 6. Assign the train combination with the highest choice probability to a passenger.

Among the multiple train combinations, the train combination with the highest choice probability is assigned to a passenger. The train choice probability is estimated by multiplying the probability of each travel time attribute. The calculation is based on the multiplication rule probability. If the passenger has an alternative route with transfers, the choice probability of transfer is multiplied as a transfer penalty. If not, the train choice probability is estimated with the choice probability of access time and egress time. The mathematical expression of estimating the train choice probability is shown in the following equation:

$$\begin{aligned} &v^* = v, \\ &s.t. pr^v = \max\left(pr^1, pr^2, pr^3, \ldots pr^v, \ \ldots, pr^w\right). \end{aligned} \tag{21}$$

Step 7. Go to Step 4 to estimate the next passenger's travel train combination until no remains.

The steps from 4 to 7 operate iteratively until estimating all passengers' train choices, since the proposed algorithm estimates the train choice for each passenger.

3.4. Performance Measure for Validating Train Choice.

The performance measures, e.g., precision, recall, accuracy, and F1 score, were used to validate the model performance. The precision, recall, accuracy, and F1 score are well-known measures for validating the performance of the model in each passenger. The values of performance measures were estimated by comparing the passenger's explored route from the assigned train combination and the actual route recorded in smart card data. Precision is defined as the accuracy of estimating true positives from the true negatives and false positives, as in equation (22). The recall is the number of true positives among the true negatives and false positives as in equation (23). The accuracy is the number of true positives and true negatives among all the passengers, as in equation (24). The F1 score is the trade-off between recall and precision, and has equal importance as in equation (25):

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{22}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{23}$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{FP} + \text{TN}}, \tag{24}$$

Table 3: Definition of notations.

Choice set generation algorithm
N: the set of the subway station number
$\mathbf{p}_i$: the vector of the travel attributes of the passenger i
t_i^{in}: the tap-in time of the passenger i
t_i^{in}: the tap-out time of the passenger i
o_i: the origin station of the passenger i
d_i: the destination station of the passenger i
$\mathbf{r}$: the vector of the attributes of the train combination
δ: the train departure time at the origin station
tr_{in}^k: the arrival time of the train for the previous segment (before transfer) at the transfer station k
tr_{out}^k: the departure time of the train for the next segment (after transfer) at the transfer station k
α: the train arrival time at the destination station
o: the origin station of the train combination
d: the destination station of the train combination
$R_{OD}(\mathbf{p}_i)$: the set of the alternative train combination for the passenger i

ECDF
$f_j(x)$: ECDF of the attribute x

Train choice algorithm
Choice set-related notations
U: the set of the passengers who have only one alternative train combination
M: the set of the passengers who have more than two alternative train combinations
P: the set of the passengers
$n(R_{OD}(\mathbf{p}_i))$: the number of the alternative train combination of passenger i
Travel time attribute-related notations
a_i: the access time of passenger i
e_i: the egress time of passenger i
tr_i: the transfer time of passenger i
a_i^o: the access time at the origin station o
e_i^d: the egress time at the destination station d
tr_i^k: the transfer time at the transfer station k
ma_u^o: the mode value of the access time of the passenger u
me_u^d: the mode value of the egress time of the passenger u
mtr_u^k: the mode value of the transfer time at the transfer station k of the passenger u
ECDF-related notations
F_a^o: the ECDF of the access time at the origin station o
F_e^d: the ECDF of the egress time at the destination station d
F_{tr}^k: the ECDF of the access time at the origin station k
Choice probability-related notations
pr: the choice probability of the train combination
pr_a: the probability of access time of the alternative train combination
pr_e: the probability of egress time of the alternative train combination
pr_{tr}: the probability of transfer time of the alternative train combination
v^*: the number of the train combination with the highest choice probability
w: the number of the alternative train combination
pr^v: the choice probability of alternative train combination v

$$F1\ \text{score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{25}$$

where TP is the true positives, FP is the false positives, TN is the true negatives, and FN is the false negatives.

4. Application

4.1. Validation of the Travel Route Estimation Results. The results of estimated travel routes and train combinations for individual passengers are validated with smart card data obtained from two private lines, i.e., Line 9 and the Shinbundang Line. The route information of passengers who get in or get off the private lines as part of their travel routes could be easily produced since the private lines facilitate transfer gates at their transfer stations. The results of the travel route estimation are compared with the actual route of trips recorded in smart card data. For example, O-D pair in Figure 3 was selected to illustrate the process of the train choice estimation. Figure 3 shows the route of the Seoul National University of Education (SNUE) Station to Dangsan Station. There are two alternative routes between SNEU Station and Dangsan Station: no-transfer route and one-transfer route. Route 1 directly connects O-D stations with no transfers, and route 2 contains one transfer at Express Terminal Station on their route. Route 1 is the no-transfer route, which is on a single line. Route 2 is a one-

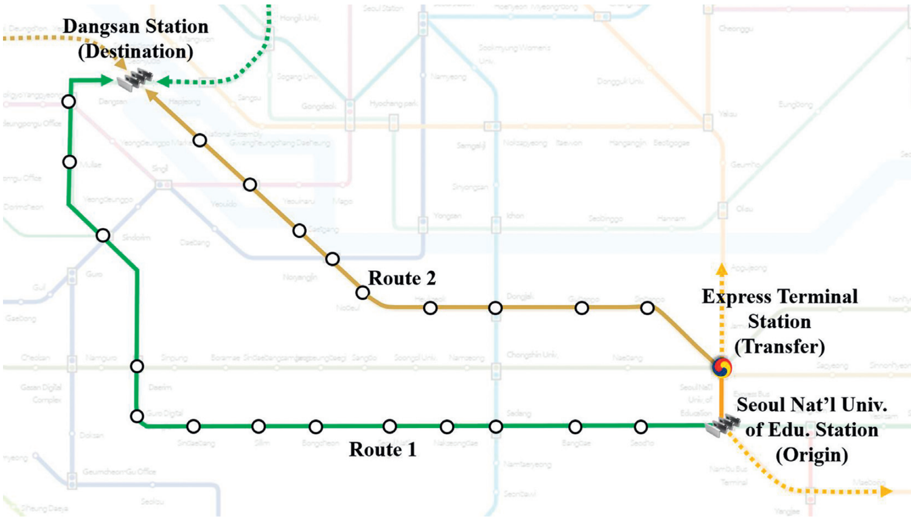


Figure 3: Illustration network with alternative routes from SNUE Station to Dangsan Station.

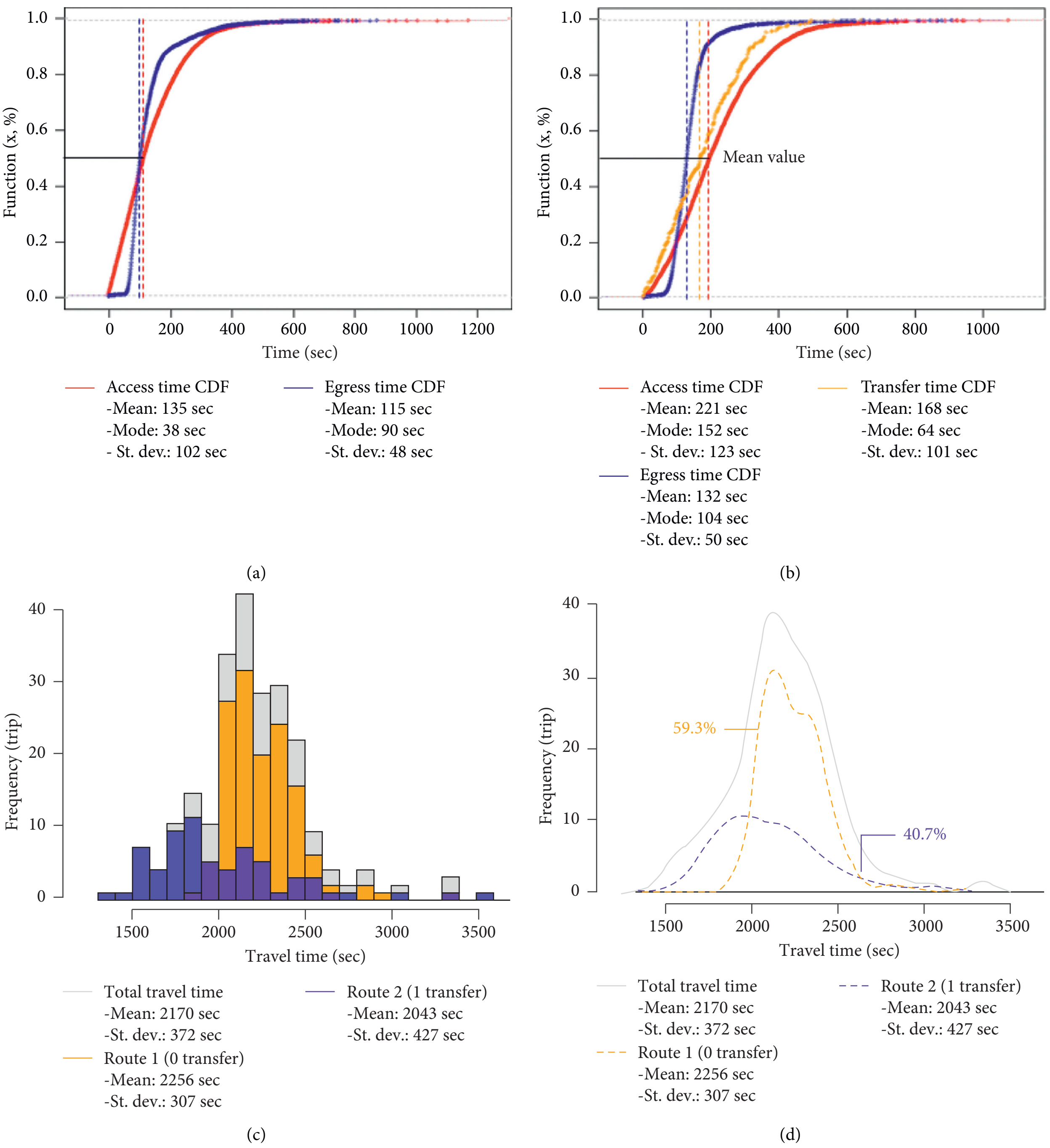
transfer route, where the Express Terminal Station connects the two lines. All ECDFs for each direction of origin station, destination station, and transfer stations were used to select the appropriate travel train combination. The alternative routes from SNUE Station to Dangsan Station are shown in Figure 3.

Figures 4(a) and 4(b) illustrate the cumulative distribution of travel time attributes, which are access time, egress time, and transfer time of routes 1 and 2.

As a result of the developed distributions, the mean of the access time of route 1 was estimated to be 135 seconds. The mode of egress time of route 1 was also estimated to be 38 seconds, and the standard deviation was 102 seconds. The mean, mode, and standard deviation of the egress time of route 1 were estimated to be 115, 90, and 48 seconds, respectively. For route 2, the average of access time, egress time, and transfer time was estimated to be 221, 132, and 168 seconds, respectively. The mode value of access time, egress time, and transfer time of route 2 was estimated to be 152, 104, and 64 seconds, respectively. The standard deviations of access time, egress time, and transfer time were estimated to be 123, 50, and 101 seconds, respectively. Figures 4(c) and 4(d) show the travel time distributions of the two routes. The grey histogram in Figure 4(c) and the grey line in Figure 4(d) represent the total travel time distribution of passengers from SNUE Station to Dangsan Station. This total travel time distribution is shown as the mixed distribution of two routes' travel time. With the distributions of access time, egress time, and transfer time, the total travel time distribution was decomposed by two distributions of respective routes. The results of the decomposed distributions are colored yellow for route 1 and blue for route 2. The mean of total travel time of OD is 2,170 seconds, and the standard deviation is 372 seconds. For route 1, the average travel time is estimated to be 2,256 seconds and the standard deviation is 307 seconds. Route 2 has 2,043 seconds for the average travel time and 427 seconds for the standard deviation of travel time. The result of the travel route estimation from SNUE Station to Dangsan Station is shown in Figure 4.

The comparison analysis was conducted to evaluate the performance of the proposed model. Three comparison models were used to compare with the proposed model. Three comparison models consist of the Gaussian mixture model (GMM) [17], maximum route length model (MRL) [9], and parametric distribution model (PDM) [20]. GMM decomposed the travel time distribution into the number of routes, assuming the Gaussian distribution. GMM assigned the train combination to a passenger with the probability distribution of each route travel time. MRL assigned the train combination to a passenger with the maximum route length (time duration) that fits within the tap-in and tap-out time of the journey. PDM assigned the train combination to a passenger based on the travel time attribute distributions, e.g., access, egress, transfer, and in-vehicle time. The access, egress, and transfer time were assumed to be gamma distribution. The waiting time and in-vehicle time were assumed to be the Poisson and uniform distributions, respectively. Each parameter of distribution was estimated to explore the passengers' route choice preference. Overall, four models, including the proposed model, were compared to evaluate the model performance.

As a result of the comparison analysis, the choice probability of route 1 was estimated to be 54.4% to 64.8%. Among the four models, the proposed model had the most similar probability at 59.3% compared with the actual route choice probability. Regarding individual train combination choice, the F1 scores of GMM, MRL, PDM, and proposed model were estimated to be 0.688, 0.739, 0.918, and 0.963, respectively. Overall, the proposed model showed the highest performance in both aggregated probabilities, such as choice probability and individual choice estimation. PDM also showed good performance with 0.918 F1 score. However, the F1 score of PDM was estimated to be lower than that of the proposed model since the errors due to the assumption of



(a)

(b)

(c)

(d)

FIGURE 4: Estimation results of SNUE Station to Dangsan Station trips. (a) Cumulative distribution for no-transfer route. (b) Cumulative distributions for one-transfer route. (c) Histogram of travel time. (d) Distribution of travel time.

distribution are involved. Especially, the assumption of uniform distribution had the greatest influence on the inaccuracy. These results implied that the proposed model estimates passengers' train choice preference more accurately than the GMM, MRL, and PDM. The travel route estimation result of the comparison models is shown in Table 4.

The results of the proposed algorithm are validated using the trips made through the private lines. As mentioned before, smart card data from the private lines provide transfer information and make it possible to identify the passenger's travel route.

From smart card data, the number of trips on private lines was counted as 472,436 trips per day. The numbers of no-transfer, one-transfer, two-transfer, and three-transfer trips are counted as 220,239, 241,114, 10,738, and 345, respectively. Table 5 shows the validation results of the travel route estimation of the proposed algorithm compared with the counted number of passengers who get in or get out of the private lines, Line 9 and Shinbundang Line, during their journey. The results

Table 4: Travel route estimation result of the comparison models.

Division	Estimated number of trips		Estimated choice probability (%)		F1 score
	Route 1	Route 2	Route 1	Route 2	
Actual	145	108	57.3	42.7	—
GMM	164	89	64.8	35.2	0.688
MRL	138	115	54.5	45.5	0.739
PA	157	96	62.1	37.9	0.918
Proposed	150	103	59.3	40.7	0.963

GMM: Gaussian mixture model. MRL: maximum route length model. PA: parametric distribution model

of no-transfer trips estimated by the proposed algorithm showed 99.7% of accuracy. For the one-transfer trips, 223,117 trips of 241,114 trips were estimated correctly, and the accuracy was estimated to be 92.5%. As a result of the two- and three-transfer trips, the accuracy was declined to be 81.1% and 71.6%, respectively. Taken together, the accuracy of the estimation result for the total trips was estimated to be 95.6%. Since the number of no-transfer and one-transfer trips accounts for 97.6% of the total validation trip samples, the estimation accuracy of the trips was estimated to be high enough to apply the proposed algorithm to the Seoul subway networks. The result of the travel route estimation is shown in Table 5.

4.2. Travel Route Estimation for Subway Network in Seoul. The travel trains for 6,313,176 daily trips were estimated to identify the route choice preference using the proposed algorithm. As results, the numbers of no-transfer, one-transfer, two-transfer, three-transfer, and four-transfer trips were estimated to be 3,402,763; 2,382,288; 411,475; 91,554; and 25,096 trips, respectively. Regarding the trip ratios of total trips, no-transfer, one-transfer, two-transfer, three-transfer, and four-transfer trips were estimated to be 53.9%, 37.7%, 6.5%, 1.5%, and 0.4%, respectively. The trip ratios of peak and nonpeak hours show similar patterns. The results of the travel route estimation on the whole network in Seoul are shown in Table 6 and Figure 5.

4.3. Evaluating the Efficiency of Subway Lines in Seoul Using the Proposed Algorithm. The proposed algorithm was applied to evaluate the efficiency of 11 subway lines on the Seoul subway network. The algorithm can produce the passenger kilometer metric for evaluating the transport efficiency of 11 lines. The Seoul Transportation Corporation (STC) has been trying to aggregate link trips using smart card data since those are the basic statistics to operate the subway network. STC roughly calculated the passenger kilometer by assigning the passenger to the shortest path because smart card data do not provide travel route information. Regarding this practical need, the travel route estimation could provide useful statistics such as passenger kilometer. The results of the travel route estimation in this study were used to measure the passenger kilometer of 11 subway lines in Seoul.

Table 5: Result of travel route estimation with private subway lines.

Division	Precision	Recall	Accuracy	F1 score
No-transfer trips	0.997	1.000	0.997	0.998
One-transfer trips	0.947	0.962	0.925	0.954
Two-transfer trips	0.832	0.946	0.811	0.885
Three-transfer trips	0.789	0.833	0.716	0.811
Total	0.968	0.979	0.956	0.974

The most widely used metric to measure transport efficiency is the value of passenger kilometer [24, 25]. Passenger kilometer is calculated by multiplying the number of passengers by the travel distance. The mathematical expression of the passenger kilometer is shown in the following equation:

$$\text{pkm} = \sum_{g}^{G} \text{tpc}_g \times \text{tdc}_g, \tag{26}$$

where pkm is the passenger kilometer value, i is the travel route ($G = 1, 2, 3, \ldots, g$), tpc is the number of passengers who traveled with the route g, and tdc is the distance of the route g (km).

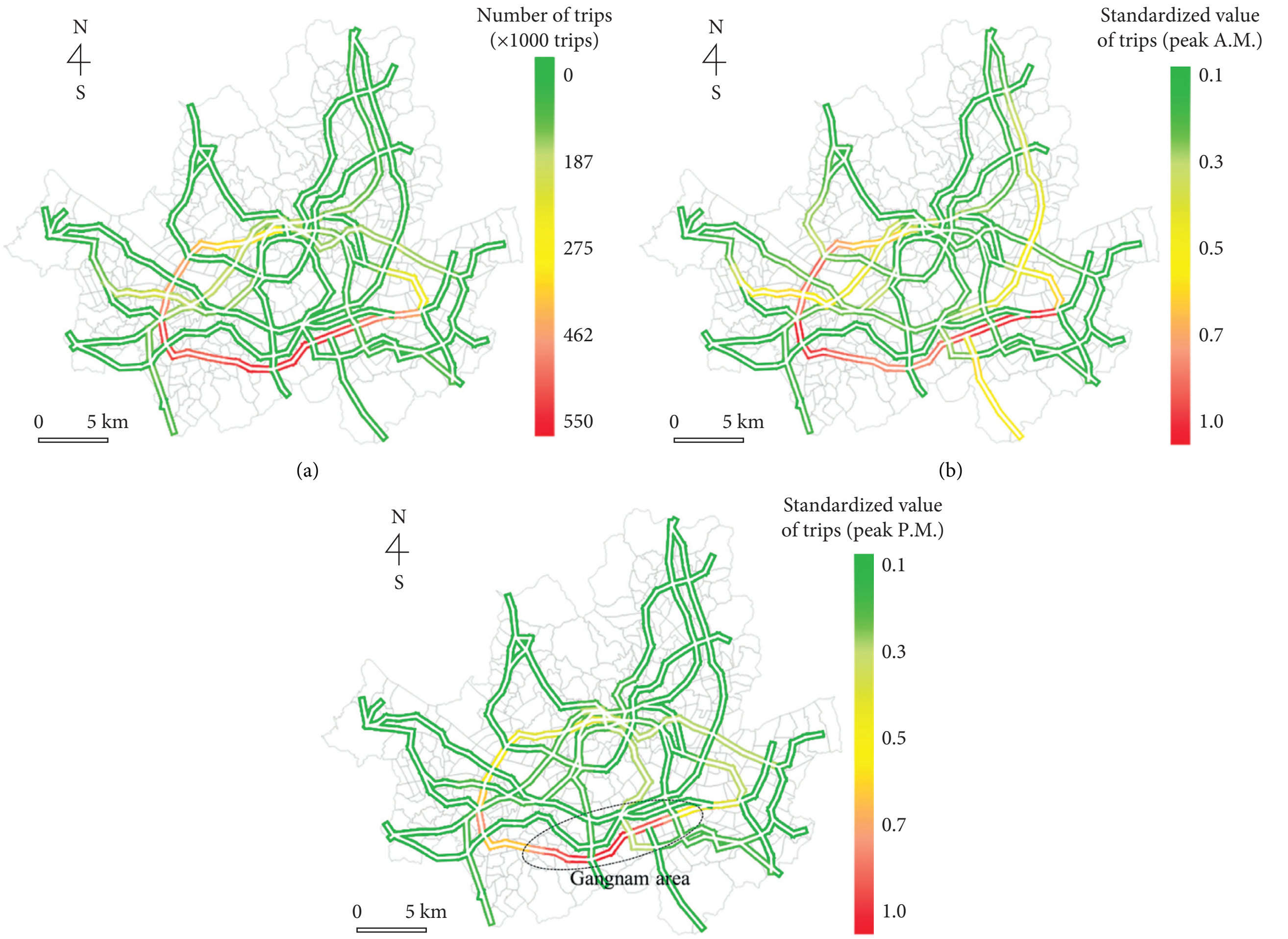
As a result of the passenger kilometer analysis, the passenger kilometer of STC was estimated to be 78,194 million passenger kilometer, and the passenger kilometer of the proposed algorithm was estimated to be 88,314 million passenger kilometer. Since the STC assigned the passenger to the shortest path, the passenger kilometer of the proposed algorithm was estimated to be about 13% higher than that of STC.

The passenger kilometer and the number of passengers were calculated by 11 subway lines. The result of the passenger kilometer of Line 2 was estimated to be 27,002 million passenger km, which is the highest value among the 11 lines. The lowest value was 1,553 million passenger kilometer, of Line 8. Since Line 2 goes through the major commercial and business areas of central Seoul, the passenger kilometer of Line 2 was estimated to be the highest among the 11 lines. For Line 8, the passenger kilometer was estimated to be the lowest because there are only 16 stations along the line and Line 8 serves on the outskirts of Seoul.

Regarding the passenger kilometer per service distance, the efficiencies of 11 lines are evaluated in the order of Line 2, Line 3, and Line 7. The efficiency order based on the number of passengers per service distance is somewhat different from that of the passenger kilometer unit. The efficiency of 11 lines based on the number of passenger units is evaluated in the order of Line 2, Line 5, and Line 7. The evaluation results of 11 lines based on two metrics are presented in Table 7.

Table 6: Results of travel route estimation for urban subway network in Seoul.

Division	Total trips (trip ratio, %)	Peak hour trips (trip ratio, %)		Nonpeak hour trips (trip ratio, %)
		AM (7:00~9:00)	PM (18:00~20:00)	
No-transfer trips	3,402,763 (53.9)	563,952 (54.1)	513,662 (55.6)	2,325,149 (53.5)
One-transfer trips	2,382,288 (37.7)	386,933 (37.1)	337,247 (36.5)	1,658,108 (38.2)
Two-transfer trips	411,475 (6.5)	70,884 (6.8)	57,512 (6.2)	283,079 (6.5)
Three-transfer trips	91,554 (1.5)	15,753 (1.5)	12,557 (1.4)	63,244 (1.5)
Four-transfer trips	25,096 (0.4)	5,189 (0.5)	3,305 (0.4)	16,602 (0.4)
Total	6,313,176 (100.0)	1,042,711 (100.0)	924,283 (100.0)	4,346,182 (100.0)



(a)

(b)

(c)

Figure 5: Visualization of estimated link trips of subway network in Seoul. (a) The number of link trips for a day. (b) Link trip density at peak A.M. (c) Link trip density at peak P.M.

Table 7: Results of passenger kilometer for subway lines in Seoul.

Subway lines	Service distance (km)		Number of passengers (trips)			Passenger kilometer (million km)		
	Distance (A)	Rank	Total (B)	Trips/service dist. (B/A)	Rank	Total (C)	Passenger kilometer/service dist. (C/A)	Rank
Line 1	195	1	7,754,053	39,764	9	12,237	63	10
Line 2	**57**	**3**	23,686,939	415,560	1	27,002	474	1
Line 3	55	5	8,027,244	145,950	4	9,523	173	2
Line 4	68	2	5,783,969	85,058	7	6,813	100	8
Line 5	50	8	7,989,509	159,790	**2**	8,105	162	4
Line 6	34	9	4,269,248	125,566	5	3,780	111	5
Line 7	56	4	8,714,031	155,608	3	9,531	170	3
Line 8	17	11	1,477,962	86,939	6	1,553	91	9
Line 9	51	7	3,650,051	71,570	8	5,152	101	7
Bundang Line	53	6	1,451,587	50,055	11	2,900	100	11
Shinbundang Line	31	10	1,015,097	35,003	10	3,168	109	6

Total trips of the Seoul subway network: 6,313,176 trips/day. Estimated passenger kilometer of Seoul network: STC, 78,194 m-pkm (100%); proposed algorithm, 88,314 m-pkm (113%).

5. Conclusion

This study proposed the travel route estimation algorithm using smart card data and train log data. The process of travel route estimation consisted of three stages: (1) generation of the train choice combinations, (2) calculation of passenger travel time attributes, and (3) development of ECDFs. The algorithm was proposed to estimate train choice for an individual subway passenger. The alternative train choice combination was generated using the passenger tap-in time and tap-out time of smart card data, and train arrival time of train log data. The travel time attributes of the passenger were calculated by each alternative train combination. The ECDFs of each type of travel time, i.e., access time, egress time, transfer time, were developed with the trip information that could only be traveled by a single train set. These developed ECDFs were used to estimate the travel route for passengers who have several alternative train combinations. The travel route was deduced by an estimated train combination with the highest probability among the alternative train combinations. The analysis is performed in two stages, i.e., validation with private subway lines and application to the entire subway network in Seoul. For the first stage, the smart card data of the private subway lines were employed to validate the results of the estimated travel train combination, since it has the exact information about the travel route transaction. For the second stage, the proposed algorithm is then applied to estimate the travel train combinations of all subway passengers on the entire subway network in Seoul.

As a result of the comparison analysis, the F1 scores of GMM, MRL, PA, and proposed model were estimated to be 0.688, 0.739, 0.918, and 0.963, respectively. This result implied that the proposed model based on ECDF estimated passengers' choice behavior more accurately than the parametric, nonparametric, and rule-based models. In particular, the proposed model could have strengths in complex subway networks such as many lines, stations, and short headways. As a result of the validation, the accuracy for the no-transfer trips, one-transfer trips, two-transfer trips, and three-transfer trips is estimated to be 99.7%, 95.1%, 84.2%, and 71.2%, respectively. The result of total trips is about 96.9%, which is reasonable to analyze the whole subway network. As a result of the travel route estimation of the whole network in Seoul, the trip ratio for no-transfer, one-transfer, two-transfer, three-transfer, and four-transfer trips was estimated to be 53.9%, 37.7%, 6.5%, 1.5%, and 0.4%, respectively. Regarding the practical application, the passenger kilometers by lines were estimated with the travel route estimation of the whole network. As a result of the passenger kilometer calculation, the passenger kilometer of the proposed algorithm was estimated to be 88,314 million passenger kilometer. Since the STC assigned the passenger to the shortest path, the passenger kilometer of the proposed algorithm was estimated to be about 13% higher than that of STC. Among the 11 subway lines, the passenger kilometer of Line 2 showed the highest value of 27,002 million passenger kilometer.

There are three main contributions to this study. First, the empirical distributions of the travel time attributes, i.e., access time, egress time, transfer time, and in-vehicle time, were developed using smart card data and train log data. Specifically, the subway station's walking characteristics were reflected on access time and egress time without assuming a specific distribution form, i.e., the Poisson and uniform distribution. Second, the real data of passengers' travel routes were used to validate the proposed method. This revealed route information (transfer gate) data provided that the proposed method showed notable accuracy in estimating the travel route of subway passengers. Third, the practical application was performed by estimating whole passengers' travel routes. The results of the efficiency evaluation of each subway line implied that passengers do not always prefer the shortest route.

The results of this paper help subway operators manage in-train and route congestion. The results also contribute to an in-depth investigation of route choice behaviors by quantifying the penalty factors on routes: transfer time and distance, access time and distance, waiting time, the number of stairs, and the congestion rate on the platform. Although we estimated the traveled trains and routes using ECDFs of time attributes, some issues remain. First, the impact of crowding and potentially being left behind needs to be considered. Second, it is required to decompose the walking time and the waiting time distribution for the access time and the transfer time. In addition, information on station amenities, such as restrooms and convenience stores, needs to be considered. Hence, our future work will incorporate crowding and facility factors to estimate the travel route of the subway passengers.

Data Availability

The data used in this research were provided by the Trlab Research Program conducted at the Seoul National University, Seoul, Republic of Korea. The data are available when readers ask the authors for academic purposes.

Conflicts of Interest

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

Authors' Contributions

Eun Hak Lee provided the software, wrote the original draft, investigated the data, visualized the data, and validated the data. Kyoungtae Kim collected data; wrote, reviewed, and edited the manuscript; and acquired funding. Seung-Young-Kho investigated the data, validated the data, and wrote, reviewed, and edited the manuscript. Dong-Kyu Kim conceptualized the data, supervised the data, designed methodology, investigated the data, involved in formal analysis, wrote, reviewed, and edited the manuscript, and acquired funding. Shin-Hyung Cho conceptualized the data, developed the methodology, investigated the data, involved in formal analysis, and wrote, reviewed, and edited the manuscript.

Acknowledgments

This research was supported by a grant from the R&D Program of the Korea Railroad Research Institute, Republic of Korea.

References

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

[2] E. H. Lee, H. Shin, S.-H. Cho, S.-Y. Kho, and D.-K. Kim, "Evaluating the efficiency of transit-oriented development using network slacks-based data envelopment analysis," *Energies*, vol. 12, no. 19, p. 3609, 2019.

[3] W. Jang, "Travel time and transfer analysis using transit smart card data," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2144, no. 1, pp. 142–149, 2010.

[4] J. Y. Park, D. J. Kim, and Y. Lim, "Use of smart card data to define public transit use in Seoul, Republic of Korea," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2063, pp. 3–9, 2008.

[5] Z. Zhao, H. N. Koutsopoulos, and J. Zhao, "Individual mobility prediction using transit smart card data," *Transportation Research Part C: Emerging Technologies*, vol. 89, pp. 19–34, 2018.

[6] E. H. Lee, H. Lee, S.-Y. Kho, and D.-K. Kim, "Evaluation of transfer efficiency between bus and subway based on data envelopment analysis using smart card data," *KSCE Journal of Civil Engineering*, vol. 23, no. 2, pp. 788–799, 2019.

[7] B. Si, M. Zhong, J. Liu, Z. Gao, and J. Wu, "Development of a transfer-cost-based logit assignment model for the Beijing rail transit network using automated fare collection data," *Journal of Advanced Transportation*, vol. 47, no. 3, pp. 297–318, 2013.

[8] X. Gong, G. Currie, Z. Liu, and X. Guo, "A disaggregate study of urban rail transit feeder transfer penalties including weather effects," *Transportation*, vol. 45, no. 5, pp. 1319–1349, 2018.

[9] E. Van Der Hurk, L. Kroon, G. Maróti, and P. Vervest, "Deduction of passengers' route choices from smart card data," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, pp. 430–440, 2014.

[10] K. K. A. Chu and R. Chapleau, "Enriching archived smart card transaction data for transit demand modeling," *Transportation Research Record*, vol. 2063, pp. 63–72, 2008.

[11] N. Nassir, M. Hickman, and Z.-L. Ma, "A strategy-based recursive path choice model for public transit smart card data," *Transportation Research Part B: Methodological*, vol. 126, pp. 528–548, 2019.

[12] J. Chan, *Rail Transit OD Matrix Estimation and Journey Time Reliability Metrics Using Automated Fare Data*, Massachusetts Institute of Technology, Cambridge, MA, USA, 2007.

[13] F. Zhou and R.-H. Xu, "Model of passenger flow assignment for Urban rail transit based on entryand exit time constraints," *Transportation Research Record*, vol. 2284, pp. 57–61, 2012.

[14] W. Zhu, H. Hu, and Z. Huang, "Calibrating rail transit assignment models with genetic algorithm and automated fare collection data," *Computer-Aided Civil and Infrastructure Engineering*, vol. 29, no. 7, pp. 518–530, 2014.

[15] L. Sun, D.-H. Lee, A. Erath, and X. Huang, "Using smart card data to extract passenger's spatio-temporal density and train's trajectory of MRT system," in *Proceedings of the . ACM SIGKDD Int. Workshop Urban Comput.*, pp. 142–148, Beijing, China, August 2012.

[16] T. Kusakabe, T. Iryo, and Y. Asakura, "Estimation method for railway passengers' train choice behavior with smart card transaction data," *Transportation*, vol. 37, no. 5, pp. 731–749, 2010.

[17] E. H. Lee, I. Lee, S.-H. Cho, S.-Y. Kho, and D.-K. Kim, "A travel behavior-based skip-stop strategy considering train choice behaviors based on smartcard data," *Sustainability*, vol. 11, no. 10, p. 2791, 2019.

[18] D. Hörcher, D. J. Graham, and R. J. Anderson, "Crowding cost estimation with large scale smart card and vehicle location data," *Transportation Research Part B: Methodological*, vol. 95, pp. 105–125, 2017.

[19] W. Li, Q. Luo, Q. Cai, and X. Zhang, "Using smart card data trimmed by train schedule to analyze metro passenger route choice with synchronous clustering," *Journal of Advanced Transportation*, vol. 2018, Article ID 2710608, 13 pages, 2018.

[20] Y. Sun and R. Xu, "Rail transit travel time reliability and estimation of passenger route choice behavior,"

Transportation Research Record: Journal of the Transportation Research Board, vol. 2275, no. 1, pp. 58–67, 2012.

[21] W. Zhu, W. Wang, and Z. Huang, “Estimating train choices of rail transit passengers with real timetable and automatic fare collection data,” *Journal of Advanced Transportation*, vol. 2017, Article ID 5824051, 12 pages, 2017.

[22] 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 ID 04015037.

[23] S. P. Hong, Y. H. Min, M. J. Park, K. M. Kim, and S. M. Oh, “Precise estimation of connections of metro passengers from Smart Card data.” *Transportation*, vol. 43, pp. 749–769, 2014.

[24] Uic Activity Report 2018, “International Union of Railway,” 2018, https://uic.org/.

[25] B. Feng, E. H. Park, H. Huang et al., “Discrete element modeling of full-scale ballasted track dynamic responses from an innovative high-speed rail testing facility,” *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2673, no. 9, pp. 107–116, 2019.

Hindawi
Journal of Advanced Transportation
Volume 2021, Article ID 6635198, 12 pages
https://doi.org/10.1155/2021/6635198

Research Article

Numerical Study on Propagative Waves in a Periodically Supported Rail Using Periodic Structure Theory

Xi Sheng,[1] **Huike Zeng**,[1] **Sara Ying Zhang**,[1] **and Ping Wang**[2]

[1]*Institute of Urban Smart Transportation and Safety Maintenance, Shenzhen University, Shenzhen 518060, China*
[2]*MOE Key Laboratory of High-Speed Railway Engineering, Southwest Jiaotong University, Chengdu 610031, China*

Correspondence should be addressed to Sara Ying Zhang; sara.zhangying@szu.edu.cn

Received 21 December 2020; Revised 22 February 2021; Accepted 31 August 2021; Published 14 October 2021

Academic Editor: Steven Chien

This paper presents the numerical study on propagative waves in a periodically supported rail below 6000 Hz. A periodic rail model, which considers the effects of both the periodic supports and the rail cross section deformation, has been established based on the periodic structure theory and the finite element method. Two selection approaches are proposed to obtain the concerned dispersion curves from the original calculation results of dispersion relations. The differences between the dispersion curves of different support conditions are studied. The propagative waves corresponding to the dispersion curves are identified by the wave modes. The influences of periodic supports on wave modes in pass bands are revealed. Further, the stop band behaviors are investigated in terms of the bounding frequencies, the standing wave characteristics, and the cross-sectional modes. The results show that eight propagative waves with distinct modes exist in a periodically supported rail below 6000 Hz. The differences between the dispersion curves of periodically and continuously supported rails are not obvious, apart from the stop band behaviors. All the bounding-frequency modes of the stop bands are associated with the standing waves. Two bounding-frequency modes of the same stop band can be regarded as two identical standing waves with the longitudinal translation of the quarter-wavelength, one of which is the so-called pinned-pinned resonance.

1. Introduction

The high-frequency rail behaviors play a significant part in the generation of railway rolling noise. It also significantly contributes to the generation of rail corrugation via the dynamic interaction with the wheels [1, 2]. Investigation of rail vibration behavior is crucial to elucidate these matters. An accurate model for the rail dynamics should take the cross-sectional deformation, waveguide structure, and periodic supports into account. In addition, the effective frequency range for the prediction should extend up to at least 5000 Hz [3].

The continuously welded rail can be seen as infinitely long. The models which truncate the rail at a particular length will artificially introduce modal behaviors. Besides, the rail is a waveguide, that is, a structure which has a uniform cross section and extends in the longitudinal direction. Thus, its motion is not composed of normal vibration modes but a series of guided waves [4]. A complex wavenumber can be used to describe the velocity and phase of the wave through the real part and the amplitude decay through the imaginary part.

Several numerical methods have been adopted to investigate the guided waves in rails. Ryue et al. [5] established a finite element (FE) model of a short length of the rail with symmetric or antisymmetric boundary conditions at both ends of its length. The modal analysis results were then used to obtain the dispersion relations, group velocities, and mode shapes. An alternative numerical method, known as the guided wave-based finite element method, has been widely employed. This method is also referred to as the semianalytical FE method [6], the waveguide FE method [7], and the wavenumber FE method [5]. In this method, the two-dimensional FE approach is used to discretize the cross section of the waveguide, while a wave solution is assumed in the longitudinal direction. Therefore, it has great advantages in modeling wave propagation of waveguides, such as rails. Initial relevant works were done by [8] to obtain the dispersion relations and cross-sectional mode shapes of propagative waves in a free undamped rail. Ryue et al. [5] investigated the waves propagating in the continuously supported rail for frequencies up to 80 kHz. Furthermore, Li et al. [9] improved the rail model by allowing multiple layers of the support to be considered in the modeling and then obtained the dispersion relations of waves. The rail supports, such as rail pads, sleepers, and ballast, are treated as a continuous layer of equivalent springs connected to the rail foot.

The straight railway track structure consists of the same units placed repeatedly along the longitudinal direction of the railway line. The perfect railway track can be seen as a one-dimensional periodic structure. Therefore, the periodic structure theory has also been introduced to investigate the wave propagation. The periodic rail support is the essential characteristic of the standard track structure, which has gained much attention. Tassilly [10] analyzed the propagation of bending waves in a periodically supported rail whose deflection was described by a differential equation of the fourth order. Predictions were made for a rail of a typical European railway track. Wang et al. [11] modeled the rail as a periodically supported Timoshenko beam considering the bending-torsion coupling. The dispersion relations of waves were calculated according to the transfer matrix method and Bloch's theorem.

One of the main effects of periodic supports is the occurrence of the pinned-pinned frequencies, where the wavelengths of different orders are closely related to the fastener spacing [12]. The pinned-pinned frequency has been associated with some forms of rail corrugation, which leads to the problems of noise and track structure damage [13]. Moreover, existing researches show that the dispersion relations of waves propagating in a periodic structure exhibit stop and pass band behaviors [14–16]. The stop bands (also called band gaps) are the frequency ranges between the dispersion curves, where wave propagation is prohibited. The other frequency ranges are called pass bands, where the dispersion curves exist and waves can propagate freely. Reference [11] used beam models to explore the stop and pass band behaviors of waves propagating in the periodically supported rail.

Although many studies as previously described focused on the wave propagation in the rail, few attempted to put insight into the effects of both the periodic supports and the rail cross-section deformation. These two factors are the key to an accurate prediction of wave propagation for rails in a real state. However, it is difficult to include the periodic support in the guided wave-based finite element method due to the model assumption, while the rail cross-section deformation cannot be considered in the periodically supported beam model. Beam models are not accurate at high frequencies since the cross section of the rail deforms significantly above 1500 Hz [17]. As the FE method has the advantage of calculating rail cross-section deformation and the periodic structure theory is particularly useful for periodically supported rails, they are both utilized to study the propagative waves in a periodically supported rail.

In this paper, a periodic rail model has been established, which considers the effects of both the periodic supports and the rail cross-section deformation. Two selection approaches are proposed to obtain the concerned dispersion curves from the original calculation results of dispersion relations. The differences between the dispersion curves of different support conditions are studied. The propagative waves corresponding to the dispersion curves are identified by the wave modes. The influences of periodic supports on wave modes in pass bands are revealed. Further, the stop band behaviors are investigated in terms of the bounding frequencies, standing wave characteristics, and cross-sectional modes.

2. Modeling of a Periodically Supported Rail

2.1. Unit Cell. The periodic rail model is established in this section, as shown in Figure 1. A short rail and a rail pad form the model which can be seen as a unit cell of the infinitely long and periodically supported rail.

A CHN60 rail considered as the isotropic elastic material is modeled with three-dimensional solid elements. The length of the model is equal to the fastener spacing. A rail pad is laid beneath the middle of the model, which is modeled with multiple discrete linear springs. The upper nodes of the springs are connected to the rail foot, while the lower nodes are fixed. These springs are modeled homogeneously in the rectangular rail pad area, whose width is equal to that of the bottom of the rail foot. The rail pad inflicts constraints on the rail in three directions, that is, vertical, longitudinal, and lateral stiffnesses. The damping is not considered in the model. The parameter values are listed in Table 1.

2.2. Periodic Boundary Condition. When an ideal elastic medium (continuous, homogeneous, isotropic, and perfectly elastic) deforms slightly, the governing equation of motion without body force can be expressed as follows [18]:

$$\rho \frac{\partial^2 \mathbf{u}(\mathbf{r},t)}{\partial t^2} = (\lambda+\mu)\nabla\nabla\cdot\mathbf{u}(\mathbf{r},t) + \mu\nabla^2\mathbf{u}(\mathbf{r},t), \qquad (1)$$

where $\boldsymbol{u}$ ($\boldsymbol{r}$, t) is the displacement field, $\boldsymbol{r}$ is the coordinate vector, t is the time, ρ is the density, λ and μ are the Lamé constants, ∇ is the Hamilton differential operator, and $\bullet$ is the inner product. According to Bloch theorem, the solution of equation (1) for a periodic structure can be expressed as follows [19]:

$$\mathbf{u}(\mathbf{r}) = \mathrm{e}^{\mathrm{i}\mathbf{K}\cdot\mathbf{r}}\mathbf{u}_{\mathbf{K}}(\mathbf{r}), \qquad (2)$$

where $\boldsymbol{K}$ is the wave vector in the reciprocal space and $\boldsymbol{u}_{\boldsymbol{K}}(\boldsymbol{r})$ is a periodical function with the same periodicity as the unit cell. The periodicity can be expressed as follows:

$$\mathbf{u}_{\mathbf{K}}(\mathbf{r}+\mathbf{a}) = \mathbf{u}_{\mathbf{K}}(\mathbf{r}), \qquad (3)$$

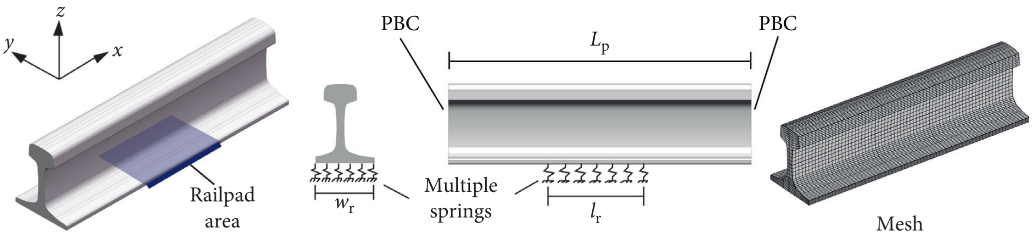


Figure 1: The periodic rail model.

Table 1: Model parameters.

Track component	Parameter	Value
CHN60 rail	Young's modulus	$E = 210$ GPa
	Poisson's ratio	$\nu = 0.3$
	Density	$\rho = 7830\,\mathrm{kg/m^3}$
	Length	$L = 0.65$ m
Rail pad	Width	$w_r = 0.15$ m
	Length	$l_r = 0.16$ m
	Vertical stiffness	$k_v = 70$ kN/mm
	Longitudinal stiffness	$k_l = 20$ kN/mm
	Lateral stiffness	$k_L = 30$ kN/mm
	Fastener spacing	$d = 0.65$ m

where $\boldsymbol{a}$ is the periodic constant vector. Substituting equation (3) into equation (2) yields the periodic boundary condition:

$$\mathbf{u}(\mathbf{r}+\mathbf{a}) = \mathrm{e}^{i\mathbf{K}\cdot(\mathbf{r}+\mathbf{a})}\mathbf{u}_{\mathbf{K}}(\mathbf{r}+\mathbf{a}) = \mathrm{e}^{i\mathbf{K}\cdot\mathbf{a}}\mathrm{e}^{i\mathbf{K}\cdot\mathbf{r}}\mathbf{u}_{\mathbf{K}}(\mathbf{r}) = \mathrm{e}^{i\mathbf{K}\cdot\mathbf{a}}\mathbf{u}(\mathbf{r}). \tag{4}$$

As the model is a one-dimensional periodic structure, the periodic constant vector $\boldsymbol{a}$, the wave vector $\boldsymbol{K}$, and the coordinate vector $\boldsymbol{r}$ can be substituted by the length of the model L, the wavenumber k, and the longitudinal coordinate x, respectively. Then, the periodic boundary condition can be written as follows:

$$\mathbf{u}(x+L) = \mathrm{e}^{ikL}\mathbf{u}(x). \tag{5}$$

The period boundary condition is set on two sides of the rail. The combination of the periodic boundary condition and the modal analysis gives the eigenvalue problem involving the dispersion relation. Then, the dispersion curves of waves propagating in a periodically supported rail can be obtained by calculating the eigenfrequencies at different wavenumbers, while the corresponding eigenvectors can be used to describe the wave modes. The dispersion curves are complex with damping considered in the model, whereas they will be real if damping is not considered. Since the wavenumber is set artificially prior to solving the eigenvalue problem, it is feasible to set the real wavenumbers rather than complex wavenumbers. Consequently, the damping is not considered in the model, and the focus of this paper is on the propagative waves, which can occur at frequencies above their "cut-on" frequencies.

2.3. Calculation of Dispersion Relations Using the Finite Element Method. To calculate the dispersion relations of propagative waves in a periodically supported rail, the finite element software is utilized to solve the model. The rail is meshed with eight-node hexahedral solid elements. In the rail pad area, each node of the bottom of the rail foot is connected by a linear three-directional spring element, which represents the constraints of the fastener. To capture the wave features at high frequencies, the in-plane mesh size of the rail cross section is shorter than 5 mm, while the mesh size in the longitudinal direction does not exceed 1 cm. The mesh sizes are proved to be fine enough in the frequency range of interest because the finer meshing gives nearly the same results. The mesh of the periodic rail model is shown in Figure 1.

By letting the wavenumber k sweep the real interval $[-4\pi/L, 6\pi/L]$ at regular intervals of $\pi/L/30$ and then calculating the corresponding eigenfrequencies, we can obtain the original dispersion relation results, as shown in Figure 2.

2.4. Selection of the Concerned Dispersion Curves. Let $f_n(k)$ denote the nth eigenfrequency when the wavenumber is equal to k. From Figure 2, we can find that the original results of dispersion relations have three features. First, the dispersion relations are symmetric with respect to $k = 0$, which means $f_n(k) = f_n(-k)$. Because the model is a periodic and symmetric structure, there exist two propagative waves, which propagate in opposite directions and have the same propagation characteristics. Second, $f_n(k)$ is a periodic function with the period $2\pi/L$. We transform equation (5) into the following:

$$\mathbf{u}(x+L) = \mathrm{e}^{ikL}\mathbf{u}(x) = \mathrm{e}^{i(kL+2\pi)}\mathbf{u}(x) = \mathrm{e}^{i(k+2\pi/L)L}\mathbf{u}(x) = \mathrm{e}^{ik*L}\mathbf{u}(x) \quad k* = k + \frac{2\pi}{L}. \tag{6}$$

From equation (6), we can find that the same modal results will be obtained if the wavenumber k is substituted by k^*. Thus, we can get $f_n(k) = f_n(k^*-)$; that is, $f_n(k) = f_n(k + 2\pi/L-)$. Third, $f_n(k)$ is symmetric with respect to

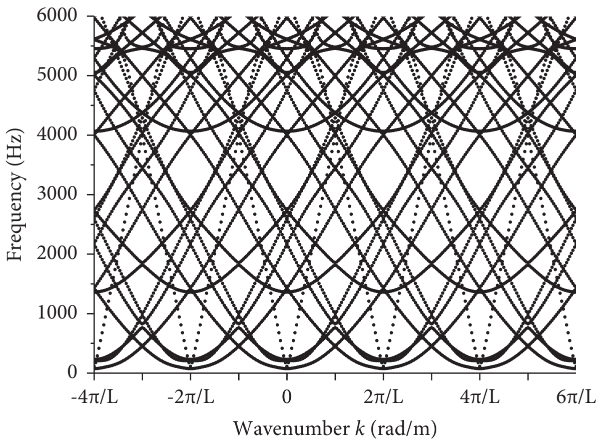


Figure 2: The original results of dispersion relations.

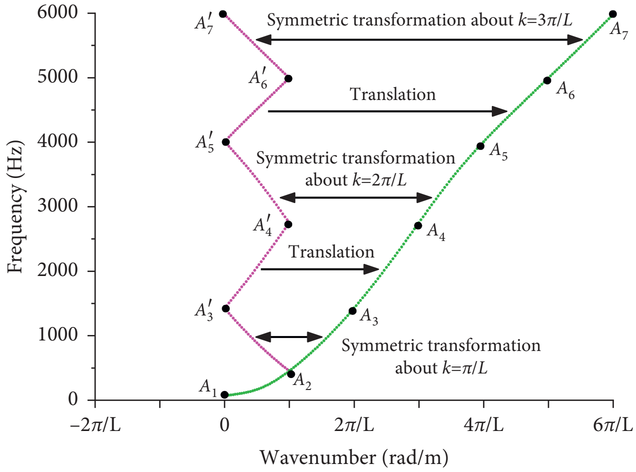


Figure 3: Selection of the dispersion curve of wave A.

$k = m\pi/L$, where m denotes the integer. This feature can be derived from the first two features, shown as follows:

$$f_n(k) = f_n(-k) \Rightarrow f_n(k) = f_n\left(\left(\pm\frac{2\pi}{L}\pm\ldots\pm\frac{2\pi}{L}\right)-k\right) \Rightarrow f_n(k) = f_n\left(m\frac{2\pi}{L}-k\right) \Rightarrow f_n(k) = f_n\left(m\frac{\pi}{L}-\left(k-m\frac{\pi}{L}\right)\right). \quad (7)$$

Based on three features, the whole dispersion relations can be obtained by the translations and symmetric transformations of the part where wavenumbers are in the interval $[0, \pi/L]$. This interval is called the irreducible Brillouin zone of the one-dimensional periodic structure. However, the dispersion relations in Figure 2 include much redundant information. Although all the points are calculated by the mathematical derivation, some dispersion curves are physically unacceptable for this model. For example, the curve where the frequency decreases with the increasing wavenumber indicates the wave having a negative group velocity; that is, the direction of the group velocity is opposite to the propagation direction. In addition, the dispersion curves of the same shape, which can be obtained by the horizontal translations of each other, indicate the same wave due to the periodicity. These phenomena arise from the mathematical mechanism, which need to be removed artificially.

Two approaches are utilized for the selection of the concerned dispersion curves. First, as previously mentioned, the dispersion curves can be obtained by the translations and symmetric transformations of the part within the irreducible Brillouin zone. Second, we can identify the wave modes of different points in the dispersion relations and then select the concerned curves from points where the corresponding wave modes are of the same type. The first approach is illustrated in Figure 3 for the case of wave A.

3. Dispersion Curves and Wave Modes in Pass Bands

3.1. Dispersion Curves. By applying the selection approaches to Figure 2, we can obtain the dispersion curves of propagative waves in a periodically supported rail below 6000 Hz, as shown in Figure 4.

Eight propagative waves (denoted by $A \sim H$) are found in a periodically supported rail below 6000 Hz. To investigate the effect of the periodic supports on the dispersion curves and simultaneously verify the precision of the model, the dispersion curves of continuously supported and free rails are calculated and shown in Figure 5. In the periodic rail model of a continuously supported rail, the rail pad area evenly covers the entire rail foot. The total stiffness of the discrete linear springs in three directions is equivalent to that of a periodically supported rail.

Below 6000 Hz, eight propagative waves are also found in continuously supported and free rails. By comparing the shapes and locations of the dispersion curves, we can find that the results for continuously supported and free rails perfectly match those of existing works done by the guided wave-based finite element method [8], which verifies the precision of the periodic rail model.

Figure 5 reveals that the overall differences between the dispersion curves of three support conditions are small, above 2000 Hz, where the wave motion in the rail has been effectively isolated from the rest of the track structure. From the comparison between the periodically supported and free rails, we can find that the periodic supports have significant effects on waves $A \sim E$ near their cut-on frequencies. However, the overall difference between the dispersion curves of periodically and continuously supported rails is not obvious below 6000 Hz. They almost coincide with each other, even near the cut-on frequencies.

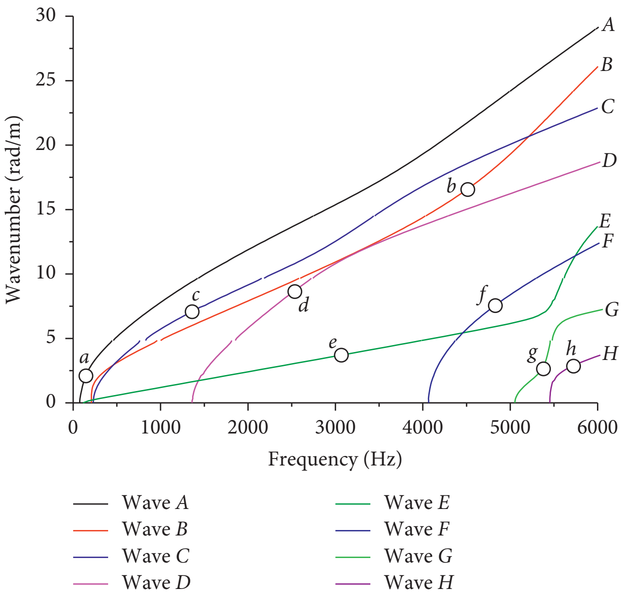


FIGURE 4: Dispersion curves of propagative waves in a periodically supported rail.

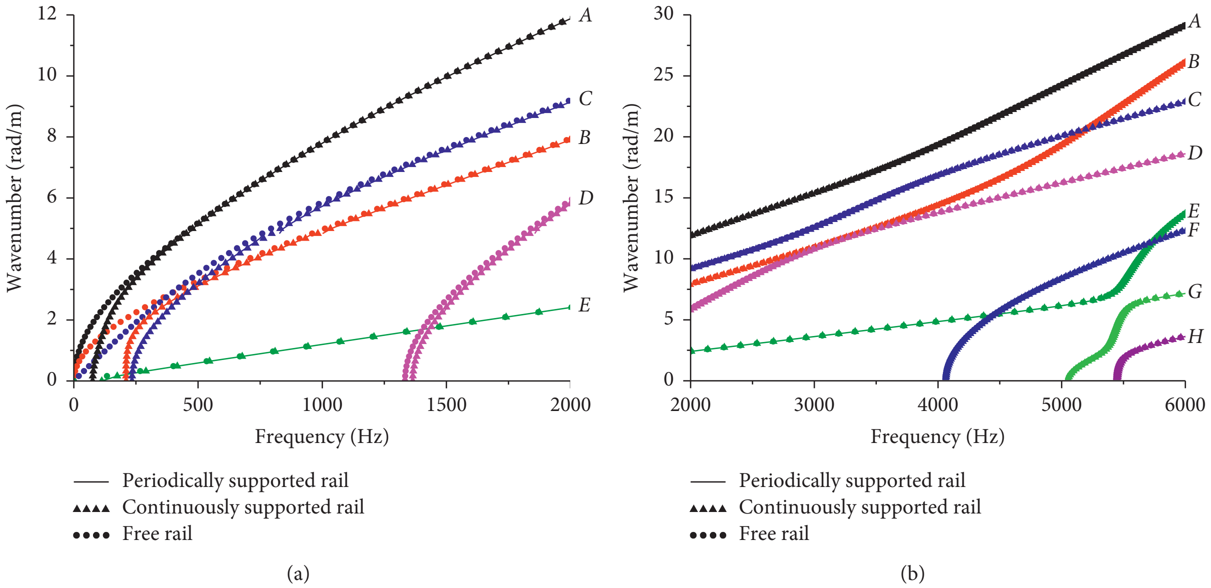


FIGURE 5: Dispersion curves of different support conditions: (a) 0–2000 Hz; (b) 2000–6000 Hz.

3.2. Wave Modes. To further characterize these waves, the wave modes corresponding to the marked points (denoted by $a\sim h$) in Figure 4 are shown in Figure 6, which contains the total displacements of the model, the out-of-plane and in-plane displacements of different cross sections, and the front view of the in-plane displacements of the cross section in the middle of the model. The colors indicate the displacement values in the particular phase state. The black lines represent the undeformed profile.

It is noteworthy that the wave modes of a wave vary with the increasing frequency. By scanning the wave modes in the whole frequency range, waves $A\sim H$ can be considered as the lateral bending wave, the vertical bending wave, the torsion wave, the bending-torsion wave, the longitudinal-vertical wave, the bending-torsion wave, the longitudinal-vertical wave, and the vertical-longitudinal wave, respectively.

As the free and continuously supported rails have uniform cross sections and boundary conditions, their modes of all cross sections are the same. The displacement amplitudes of different cross sections are identical, while the values are not same because of the phase difference. However, the periodic supports of fasteners lead to distinct boundary conditions at different positions. To compare the modes of different cross sections for a periodically supported rail, the wave modes with the increasing frequency are shown in Figure 7. The colors indicate the displacement amplitude, which is different to Figure 6.

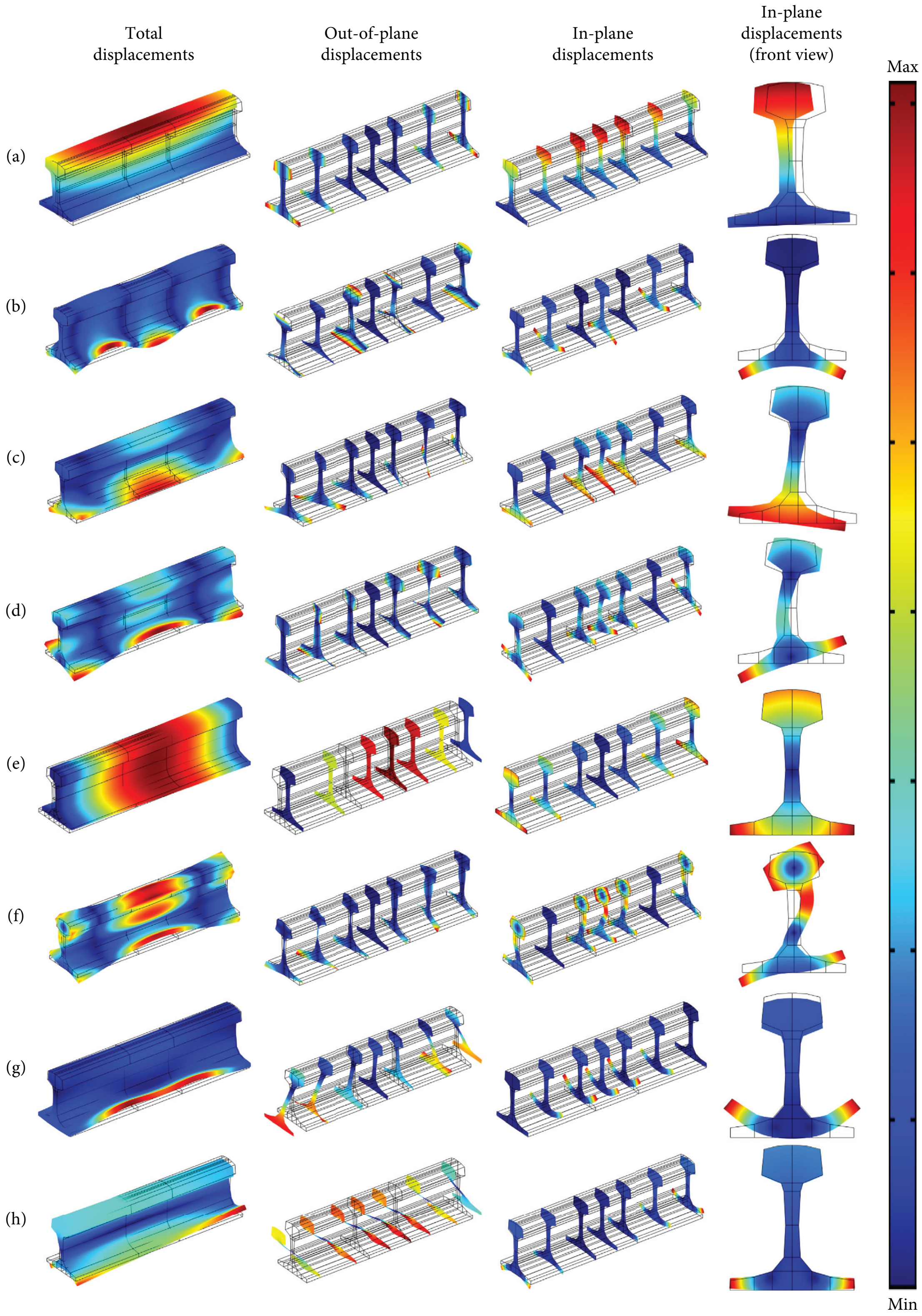


Figure 6: Wave modes corresponding to the marked points (a)~(h).

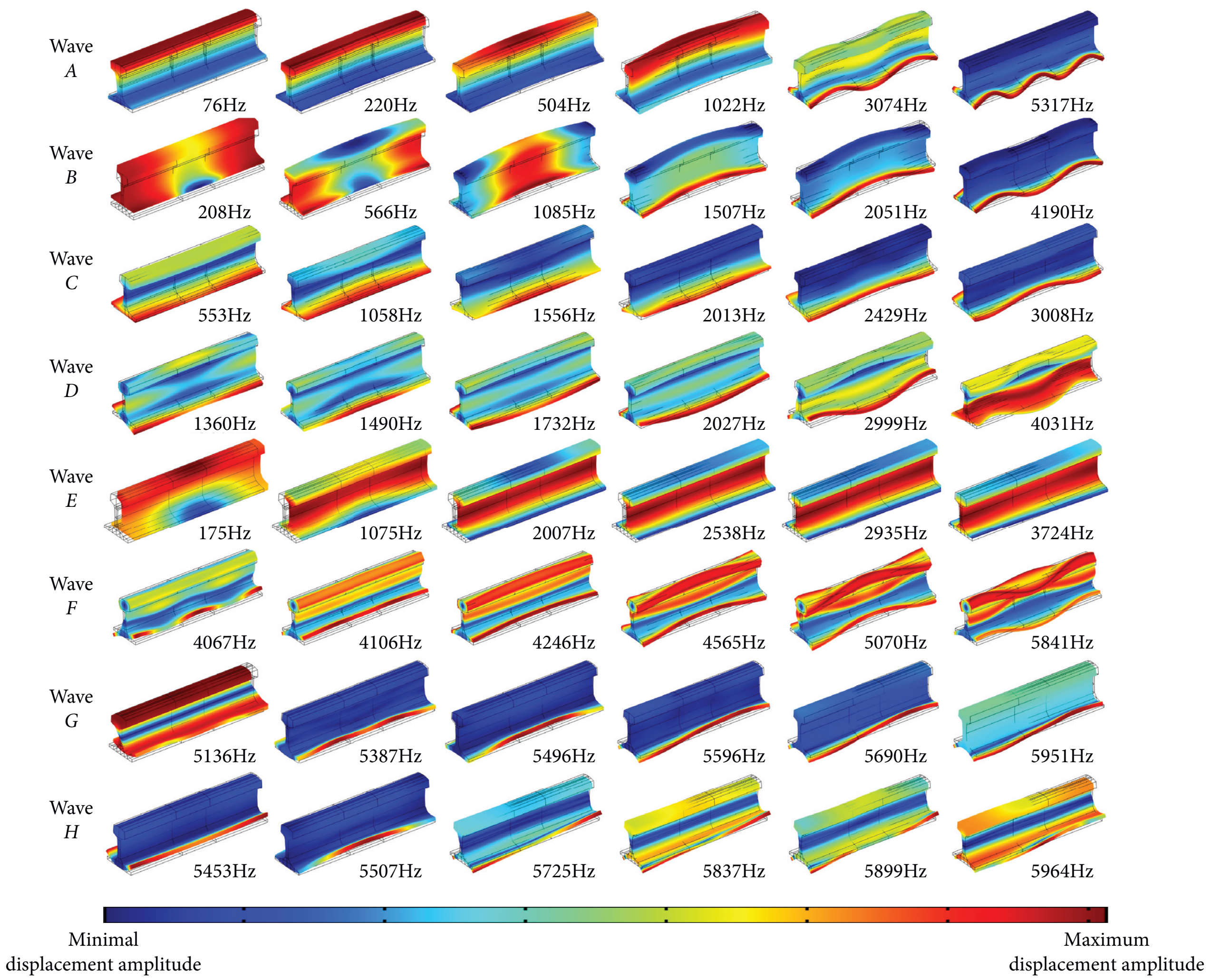


FIGURE 7: The wave modes of eight propagative waves.

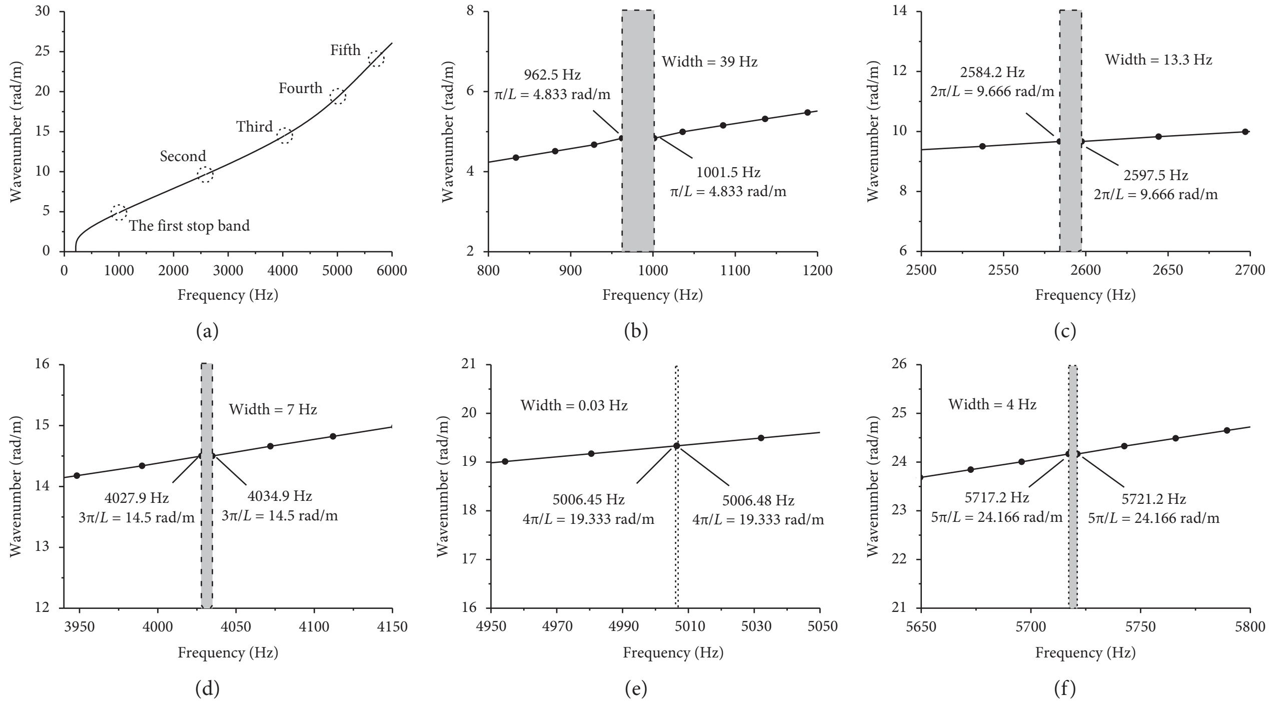


FIGURE 8: The dispersion curves and stop bands of the wave *B*: (a) the dispersion curves; (b–f) the first to fifth stop bands.

Table 2: The bounding frequencies of stop bands.

		Stop band					
		1st	2nd	3rd	4th	5th	6th
Bounding frequencies width (Hz)	*A*	445.2–459.7 14.5	1424.5–1430.2 5.7	2746.9–2750.5 3.6	4005.2–4005.4 0.2	4991.4–4996.1 4.7	5975.7–5981.1 5.4
	B	962.5–1001.5 39.0	2584.2–2597.5 13.3	4027.9–4034.9 7	5006.45–5006.48 0.03	5717.2–5721.2 4	
	C	763.0–827.7 64.7	2157.6–2183 25.4	3438.5–3449.7 11.2	4755.1–4755.3 0.2		
	D	1815.1–1855.3 40.2	2725–2741.6 16.6	4287.2–4288.8 1.6			
	E	3982.7–3985.1 2.4	5601.2–5617.3 16.1				
	F	4371.1–4379 7.9	5295.1–5298.1 3				
	G	5457.8–5480.4 22.6					
Wavenumber (rad/m)		$\pi/d = 4.833$	$2\pi/d = 9.666$	$3\pi/d = 14.500$	$4\pi/d = 19.333$	$5\pi/d = 24.166$	$6\pi/d = 28.999$
Wavelength (m)		$(2/1)\ d = 1.3$	$(2/2)\ d = 0.65$	$(2/3)\ d = 0.433$	$(2/4)\ d = 0.325$	$(2/5)\ d = 0.26$	$(2/6)\ d = 0.217$

Figure 7 shows that the periodic supports result in the cross-sectional mode differences between the rail pad and non-rail pad areas. The influences of periodic supports on eight propagative waves are as follows: (1) the modes of different cross-sections for wave A are same at low frequencies. The periodic supports have little influence on wave *A*. (2) The upper bounds of influenced frequency ranges for waves $B \sim G$ are about 1500 Hz, 2000 Hz, 1700 Hz, 2500 Hz, 4100 Hz and 5600 Hz, respectively. The influences vanish above those frequencies. (3) The cross-sectional modes of rail pad and non-rail pad areas are different below 6000 Hz for wave *H*.

4. Stop Band Behaviors

As can be seen in Figure 4, the dispersion curves of propagative waves in a periodically supported rail are discontinuous. The stop bands alternate with the pass bands. In this section, the stop band behaviors are analyzed in terms of the bounding frequencies, the standing wave characteristics, and the cross-sectional modes.

4.1. The Bounding Frequencies. As the wave *B* (vertical bending wave) is typical in the track dynamics studies, this section takes it as an illustration to elaborate the bounding-frequency properties. The dispersion curves and stop bands of the wave *B* are shown in Figure 8.

Figure 8 reveals that the dispersion relations are discontinuous at $k = n\pi/L$. No curve enters into the frequency ranges of 962.5–1001.5 Hz, 2584.2–2597.5 Hz, 4027.9–4034.9 Hz, 5006.45–5006.48 Hz, and 5717.2–5721.2 Hz. These frequency intervals are called the stop bands, and therefore, five stop bands of the wave *B* exist below 6000 Hz.

The bounding frequencies of waves $A \sim G$ are summarized in Table 2. The wave *H* has no stop band below 6000 Hz.

Below 6000 Hz, the numbers of the stop bands of waves $A \sim G$ are six, five, four, three, two, two, and one, respectively. For each wave, the wavenumber and the wavelength of the *N*th stop band are given by $k_N = N\pi/L = N\pi/d$ and $\lambda_N = 2\pi/k_N = (2/N)\ d$, respectively; the first stop band has the maximum width compared with the higher-order stop bands.

4.2. The Standing Wave Characteristics. The wave modes at the stop-band bounding frequencies of waves $A \sim G$ are shown in Figure 9. The colors indicate the displacement values in the particular phase state.

All the bounding-frequency modes of the stop bands are associated with the standing waves. The typical cross sections of these standing waves can be divided into two groups. The section group #1 includes the sections above the center of the rail pad area and at the *N*-equal-part division points of the span. The section group #2 includes the sections at the centers of *N*-equal-part segments of the span. These two section groups can be used to locate the cross-sections of nodes and antinodes.

One of two bounding-frequency modes of the same stop band is the so-called pinned-pinned resonance, that is, the modes with the red asterisks at their top-left corners in Figure 9. Either the lower-bounding-frequency mode or the upper-bounding-frequency mode corresponds to the pinned-pinned resonance, which is related to the order of the stop band. Besides, the lower-bounding-frequency mode of the first stop band of the wave *E* corresponds to the first-order longitudinal pinned-pinned resonance. The sections of group #1 have no out-of-plane displacements in this lower-bounding-frequency mode.

Furthermore, we can find that the two bounding-frequency modes of the same stop band can be regarded as two identical standing waves with the longitudinal translation of the quarter-wavelength. Thus, the minimal longitudinal distance between the cross sections of nodes (or antinodes) in two bounding-frequency modes is given by $\lambda_N/4$. Cross sections of nodes and antinodes in the lower-bounding-frequency mode coincide with those of antinodes and nodes in the upper-bounding-frequency mode, respectively.

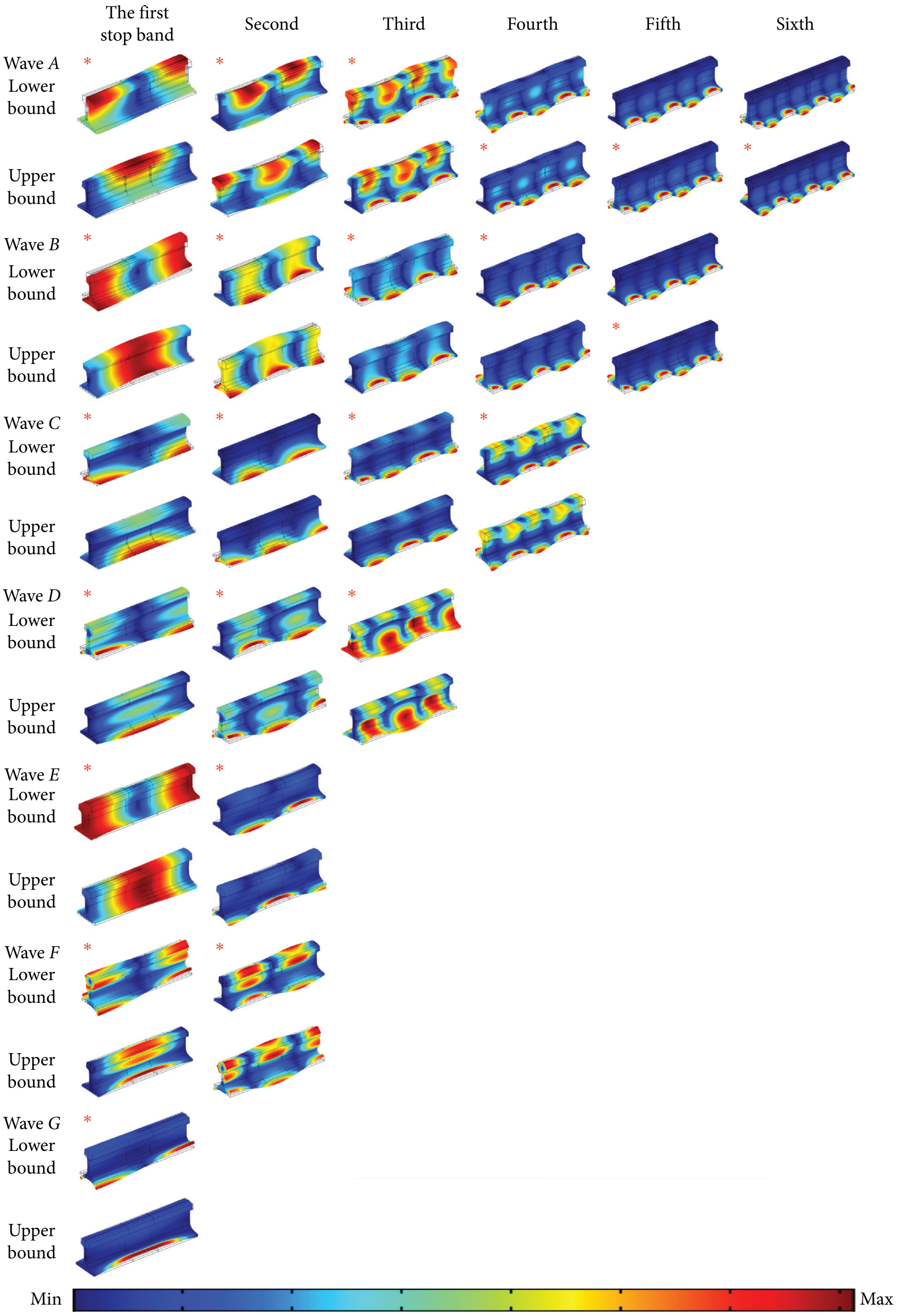


Figure 9: The wave modes at the stop-band bounding frequencies of waves *A*~*G*.

4.3. Cross-Sectional Modes. To elaborate the wave motion and the deformation at different bounding frequencies, the cross-sectional modes of the typical cross sections (i.e., section groups #1 and #2) are investigated in terms of in-plane and out-of-plane displacements in this section. The cross-sectional modes of propagative waves at the lower-bounding frequencies are shown in Table 3.

With regard to the cross sections of two section groups in a lower-bounding-frequency mode, either in-plane or out-of-plane displacements are zeros, but not both. Besides, if

Table 3: Cross-sectional modes of propagative waves at the lower-bounding frequencies.

Stop band		Section group #1			Section group #2	
		In-plane	Out-of-plane		In-plane	Out-of-plane
Wave *A*	1st~3rd	--	Torsion about the vertical axis	1st	Horizontal translation Small rotation about the longitudinal axis	--
	4th~6th	Rotation of the foot	--	2nd	Rotation of the foot Horizontal translation of the head Bending of the web	--
		Bending of the web		3rd	Same as previously mentioned (with larger amplitudes)	--
		Out-of-phase bending of the foot ends The horizontal translation of the head diminishes with the increasing frequency		4th~6th	--	Torsions of the head and web about the vertical axis Torsion of the foot about the horizontal axis; the higher-order stop band has the larger amplitude
Wave *B*	1st	--	Rotation about the horizontal axis	1st	Vertical translation Negligible foot flapping	--
	2nd~4th	--	S-shape deformations	2nd~4th	With the increasing frequency, the foot flapping becomes dominant, and the vertical translation of the head diminishes	--
	5th	Vertical deformation of the web Significant foot flapping	--	5th	--	S-shape deformations
Wave *C*	1st	--	Torsion about the vertical axis	1st	Rotation of the cross section Small bending of the web	--
	2nd	--	Torsion of the foot about the vertical axis	2nd	Bending of the web Rotation of the foot	--
	3rd~4th	--	Torsion of the head about the vertical axis Torsion of the foot about the horizontal axis	3rd~4th	High-order bending of the web Out-of-phase bending of the foot ends Horizontal translation of the head The center of the foot is immobile	--
D	1st~3rd	--	Torsion about the vertical axis; the higher-order stop band has the larger amplitude	1st~2nd	Bending of the web	--
				3rd	Bending of the web Rotations of the head and the foot Horizontal translation of the foot	--
E	1st	Tension and compression of the web Foot flapping	--	1st	--	Longitudinal translation Small bending of the cross section
	2nd	--	Longitudinal translation Bending of the cross section	2nd	Tension and compression of the web Foot flapping The center of the foot is immobile	--
F	1st~2nd	--	Torsion about the vertical axis	1st~2nd	High-order bending of the web Out-of-phase bending of the foot ends Rotation of the head The center of the head is immobile	--
G	1st	--	S-shape deformations	1st	Tension and compression of the web Foot flapping	--

one section group has no in-plane displacement, the other section group must have no out-of-plane displacement, and vice versa. For waves *A*, *B*, and *E*, the sections without in-plane or out-of-plane displacements may come from either section group, which is determined by the order of the stop band.

As previously mentioned, the two bounding-frequency modes of the same stop band can be regarded as two identical standing waves with the longitudinal translation of the quarter-wavelength. For each stop band of waves, the cross-sectional modes of section group #1 at lower and upper boundary frequencies are identical to those of group #2 at upper and lower boundary frequencies, respectively. It has been verified by investigating the cross-sectional modes of propagative waves at the upper-bounding frequencies.

5. Conclusions

The main conclusions can be drawn as follows:

(1) Eight propagative waves with distinct modes exist in a periodically supported rail below 6000 Hz. The overall differences between the dispersion curves of three support conditions (periodic supports, continuous supports, and no supports) are small, above 2000 Hz. The differences between the dispersion curves of periodically and continuously supported rails are not obvious, apart from the stop band behaviors. However, the periodic supports result in the cross-sectional mode differences between the rail pad and non-rail pad areas.

(2) The stop-band numbers of eight propagative waves are six, five, four, three, two, two, one, and zero, respectively. For each wave, the wavenumber and the wavelength of the *N*th stop band are given by $N\pi/d$ and $(2/N)\ d$, where d is the fastener spacing; the first stop band has the maximum width compared with the higher-order stop bands.

(3) All the bounding-frequency modes of the stop bands are associated with the standing waves. Two bounding-frequency modes of the same stop band can be regarded as two identical standing waves with the longitudinal translation of the quarter-wavelength, one of which is the so-called pinned-pinned resonance.

(4) With regard to typical standing-wave sections in bounding-frequency modes, either in-plane or out-of-plane displacements are zeros, but not both. These typical standing-wave sections can be divided into two groups. If one section group has no in-plane displacement, the other section group must have no out-of-plane displacement, and vice versa.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Acknowledgments

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: the National Natural Science Foundation of China (no. 52008264), Special Fund for Science and Technology Innovation Strategy of Guangdong Province (2020A0505100065), and Shenzhen Science and Technology Program (KQTD20180412181337494).

References

[1] K. Wei, Y. Dou, F. Wang, P. Niu, P. Wang, and Z. Luo, "High-frequency random vibration analysis of a high-speed vehicle-track system with the frequency-dependent dynamic properties of rail pads using a hybrid SEM-SM method," *Vehicle System Dynamics*, vol. 56, no. 12, pp. 1838–1863, 2018.

[2] W. Zhai, P. Liu, J. Lin, and K. Wang, "Experimental investigation on vibration behaviour of a CRH train at speed of 350 km/h," *International Journal of Reality Therapy*, vol. 3, no. 1, pp. 1–16, 2015.

[3] D. J. Thompson, *Railway Noise and Vibration: Mechanisms, Modelling and Means of Control*, Elsevier, Oxford, UK, 2008.

[4] M. Yuan, P. W. Tse, W. Xuan, and X. Wenjin, "Extraction of least-dispersive ultrasonic guided wave mode in rail track based on floquet-bloch theory," *Shock and Vibration*, vol. 2021, Article ID 6685450, 10 pages, 2021.

[5] J. Ryue, D. J. Thompson, P. R. White, and D. R. Thompson, "Investigations of propagating wave types in railway tracks at high frequencies," *Journal of Sound and Vibration*, vol. 315, no. 1-2, pp. 157–175, 2008.

[6] T. Hayashi, W.-J. Song, and J. L. Rose, "Guided wave dispersion curves for a bar with an arbitrary cross-section, a rod and rail example," *Ultrasonics*, vol. 41, no. 3, pp. 175–183, 2003.

[7] P. W. Loveday, "Analysis of piezoelectric ultrasonic transducers attached to waveguides using waveguide finite elements," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 54, no. 10, pp. 2045–2051, 2007.

[8] L. Gavrić, "Computation of propagative waves in free rail using a finite element technique," *Journal of Sound and Vibration*, vol. 185, no. 3, pp. 531–543, 1995.

[9] W. Li, R. A. Dwight, and T. Zhang, "On the study of vibration of a supported railway rail using the semi-analytical finite element method," *Journal of Sound and Vibration*, vol. 345, pp. 121–145, 2015.

[10] E. Tassilly, "Propagation of bending waves in a periodic beam," *International Journal of Engineering Science*, vol. 25, no. 1, pp. 85–94, 1987.

[11] P. Wang, Q. Yi, C. Zhao, and M. Xing, "Elastic wave propagation characteristics of periodic track structure in high-speed railway," *Journal of Vibration and Control*, vol. 25, no. 3, pp. 517–528, 2019.

[12] M. Oregui, Z. Li, and R. Dollevoet, "An investigation into the modeling of railway fastening," *International Journal of Mechanical Sciences*, vol. 92, pp. 1–11, 2015.

[13] S. L. Grassie, “Rail corrugation: characteristics, causes, and treatments,” *Proceedings of the Institution of Mechanical Engineers - Part F: Journal of Rail and Rapid Transit*, vol. 223, no. 6, pp. 581–596, 2009.

[14] M. S. Kushwaha, P. Halevi, L. Dobrzynski, and B. Djafari-Rouhani, “Acoustic band structure of periodic elastic composites,” *Physical Review Letters*, vol. 71, no. 13, pp. 2022–2025, 1993.

[15] X. Sheng, C.-Y. Zhao, Q. Yi, P. Wang, and M.-T. Xing, “Engineered metabarrier as shield from longitudinal waves: band gap properties and optimization mechanisms,” *Journal of Zhejiang University - Science*, vol. 19, no. 9, pp. 663–675, 2018.

[16] P. Wang, Q. Yi, C. Zhao, M. Xing, and J. Tang, “Wave propagation in periodic track structures: band-gap behaviours and formation mechanisms,” *Archive of Applied Mechanics*, vol. 87, no. 3, pp. 503–519, 2017.

[17] M. Oregui, Z. Li, and R. Dollevoet, “An investigation into the vertical dynamics of tracks with monoblock sleepers with a 3D finite-element model,” *Proceedings of the Institution of Mechanical Engineers - Part F: Journal of Rail and Rapid Transit*, vol. 230, no. 3, pp. 891–908, 2016.

[18] A. C. Eringen and E. S. Suhubi, *Elastodynamics, Volume II: Linear Theory*, Academic Press, New York, NY, USA, 1975.

[19] Y.-F. Wang, Y.-S. Wang, and X.-X. Su, “Large bandgaps of two-dimensional phononic crystals with cross-like holes,” *Journal of Applied Physics*, vol. 110, no. 11, p. 113520, 2011.

Hindawi
Journal of Advanced Transportation
Volume 2021, Article ID 6934998, 10 pages
https://doi.org/10.1155/2021/6934998

Research Article

Identifying the Service Areas and Travel Demand of the Commuter Customized Bus Based on Mobile Phone Signaling Data

Jingyuan Wang[1] **and Meng Zhang**[1,2]

[1] *College of Civil and Transportation Engineering, Shenzhen University, Shenzhen 518060, China*
[2] *Guangdong Planning and Designing Institute of Telecommunications, Guangzhou, Guangdong 510630, China*

Correspondence should be addressed to Meng Zhang; zhangmeng2018@email.szu.edu.cn

Received 24 May 2021; Revised 4 September 2021; Accepted 11 September 2021; Published 13 October 2021

Academic Editor: Xinyue Xu

In recent years, customized bus (CB), as a complementary form of urban public transport, can reduce residents' travel costs, alleviate urban traffic congestion, reduce vehicle exhaust emissions, and contribute to the sustainable development of society. At present, customized bus travel demand information collection method is passive. There exist disadvantages such as the amount of information obtained is less, the access method is relatively single, and more potential travel demands cannot be met. This study aims to combine mobile phone signaling data, point of interest (POI) data, and secondary property price data to propose a method for identifying the service areas of commuter CB and travel demand. Firstly, mobile phone signaling data is preprocessed to identify the commuter's location of employment and residence. Based on this, the time-space potential model for commuter CB is proposed. Secondly, objective factors affecting commuters' choice to take commuter CB are used as model input variables. Logistic regression models are applied to estimate the probability of the grids being used as commuter CB service areas and the probability of the existence of potential travel demand in the grids and, further, to dig into the time-space distribution characteristics of people with potential demand for CB travel and analyze the distribution of high hotspot service areas. Finally, the analysis is carried out with practical cases and three lines are used as examples. The results show that the operating companies are profitable without government subsidies, which confirms the effectiveness of the method proposed in this paper in practical applications.

1. Introduction

As a new innovative public transport mode, the CB advocates energy saving and emission reduction, green travel, alleviating urban traffic congestion, and providing people with high-quality travel services in a "point-to-point" way [1, 2]. CB originated from the idea of "car-sharing." It was introduced in 1948 by the organization "Sefage" in Sweden to save transportation costs for families who did not own a car [3]. Travel demand is an important part of customized bus route planning. Before most scholars study the route planning framework, they need to analyze the travel demand initially. K Tsubouchi et al. [4] applied the Internet and big data to develop a demand-responsive bus system that could be adapted to different city types. Qiu et al. [5] investigated a method to improve the performance of flexible route buses in an operational environment with uncertain travel demand. Scott et al. [6] researched both 'point-to-point' and 'round-trip' modes in London and predicted future demand for customized buses in London. ANand Lo [7] proposed a two-stage solution algorithm, compared to the traditional robustness formulation to determine the service with reliability using a two-stage formulation. Liu et al. [8] proposed a new commuter minibus transit system with on-demand interaction. The authors evaluated and compared the performance of CB, PC, and conventional public transportation systems through travel cost, travel time, and fuel consumption. Lyu et al. [9] proposed a CB-Planner method for a bus line planning framework with multiple travel data sources and designed a heuristic solution framework.

China's CB development started late and is still in the development stage. Zhong et al. [10] collected passenger travel demand information through online questionnaires and a mobile phone app and identified a suitable passenger flow catchment area division method. By considering the station traffic volume and regional capacity allocation, a suitable regional clustering method for passenger flow distribution is established. Cheng et al. [11] used the data from the public bus smart card to mine potential CB demand points. Yu et al. [12] planned CB stops and routes based on large amounts of demands data. Liu et al. [13] proposed a visual analysis method. They evaluated the actual, dynamically changing travel demand and planned the routes for the nighttime CB system. The reliability of the method was verified with cases.

At present, many scholars mainly research line optimization, station location, and price strategy and have achieved certain research achievements [14–16]. And the research on commuter CB travel demand is rather inadequate. Existing ways of collecting information on CB travel demand are mainly through online collection (e.g., Ma et al. proposed a framework of CB methods based on online questionnaires to obtain travel demand [17]) or through offline questionnaires in some large residential areas, commercial areas, transportation hubs, and other areas (e.g., Li et al. used RP and SP questionnaires to research the factors of influencing the potential travel demand for CB in Shanghai, China [18]). However, this passive way of collecting travel demand information is time-consuming and costly. In addition, due to the incomplete coverage and low audience level of the current CB travel demand information collection, the mining of the potential commuter CB travel demand population is neglected. Only by collecting data online or offline for a certain region, it is inevitable that the data collected for the study of travel demand is not large enough and the coverage is not extensive. There are more potential travel demands that cannot be met.

In view of the existing problems and combined with big data processing technology, this paper proposes commuter CB service areas and travel demand identification method based on mobile phone signaling data. With the following main contributions: (1) Combining mobile phone signaling data and using big data processing technology, the distribution characteristics of commuters' workplace and residence are identified. Based on the above, a time-space potential model of commuter CB travel is established and an algorithm is designed to solve it. (2) Using the unit grid as the fundamental unit, we choose the factors affecting passengers' choice of the commuter CB as the input parameters of the model. The logistic regression model is constructed and solved by SPSS software, to study the time-space distribution characteristics of people with potential commuter CB travel demand and to further identify the service areas of commuter CB and travel demand.

The rest of the paper is organized as follows. In Section 2, a brief description of the data types used in the paper is given. In Section 3, the commuter CB service areas and travel demand identification method are proposed. The central city of Chongqing, China, is used as a case study for demonstration in Section 4. The main findings of the paper are briefly summarized, and further perspectives on the following research on CB travel demand are discussed in Section 5.

2. Data Description

The data used in this paper involve three parts: mobile phone signaling data, POI data of rail stations and bus stops, and data of secondary housing prices around where commuters reside.

(i) Mobile phone signaling data: it is provided by the operator of China Unicom in Chongqing, China. It has covered 38 districts and counties in the city for mobile phone signaling monitoring, with signaling collection interval of 30–60 min. The average number of daily subscribers is 4.7 million. The average number of valid signaling data records for a single user is 26. In this paper, about 43 million data pieces of China Unicom in August 2019 are selected as the research data to identify the space-time distribution characteristics of commuters' occupation and residence. And 175,794 users from 7:00 a.m. to 9:00 a.m. on a working day in August are chosen as the research data for potential travel demand mining.

(ii) POI data of rail stations and bus stations: the POI data of the study area including 10,780 bus stations and 158 rail stations are crawled in Python programming language by retrieving the Gaode API interface. The POI attributes information included station ID, longitude, and latitude.

(iii) Secondary house prices data: by crawling the second-hand house prices on the websites of 58 TongCheng and LianJia in China, we obtain the name of each community, convert it to latitude and longitude coordinates, and obtain its spatial geographic information. The mean value of the secondary house price near the commuter's residence is used as the input parameter of the model, and this feature is used to represent the income of the commuter.

3. Identification Method of Service Areas and Travel Demand

In the process of generating mobile phone signaling data, the natural environment, interference from human factors, and other conditions can lead to error in the location of cellular cells, and there may be missing data and duplication. At first, the abnormal data are cleaned, and on this basis, the origin (O) and destination (D) of commuters in the study area are identified using the training method proposed in [19]. The characteristics of commuters' occupational and residential distribution are obtained. Based on the time-space distribution characteristics of commuter travelers' occupations and residences, a time-space potential model of commuter CB is established. We considered the influence factors as the input parameters of the model and established logistic regression model. We use the model to predict the study area and select the areas that meet the conditions as the commuter CB service area. Based on this, we further identify the potential commuter CB travel demand population.

3.1. Time-Space Potential Model. In this paper, based on mobile phone signaling data, the travel regularity of commuters, the similarity of travel time and spatial distribution of work and residence, and the possibility of taking commuter CB in time-space distribution are comprehensively considered. Based on the shared travel model framework, the distribution characteristics in two dimensions of time and space are considered, and based on the literature [20], the time-space potential model of commuter CB is proposed.

The model takes commuter travelers as the research object. We take the time difference between commuters leaving their places of residence and the distance difference between commuters' places of work and residence as independent variables. Due to the difference between time and distance units, maximum-minimum normalization is used to convert them into dimensionless expressions and introduce weighting factors. The objective function is to calculate the value of time-space potential between commuters. The model takes into account the shorter time difference between commuters in terms of travel time and the smaller distance between commuters' residence and workplace in the spatial dimension. To a certain extent, it indicates the greater potential of commuters who can travel by the same transportation mode. Therefore, when certain conditions are met, it is considered that there is a potential similar travel demand between commuters in both temporal and spatial dimensions. The formula of the model is defined as

$$\mathrm{TPV}(i,j) = \alpha\left[\frac{T(i,j)-\min(T)}{\max(T)-\min(T)}\right] + \gamma\left[\frac{S(i,j)-\min(S)}{\max(S)-\min(S)}\right] + \lambda\left[\frac{L(i,j)-\min(L)}{\max(L)-\min(L)}\right]. \tag{1}$$

Equation (1) constraint is

$$\begin{aligned} T(i,j) &= \frac{t(i,j)}{\Delta T},\\ t(i,j) &= t_i - t_j,\\ T &< \delta,\\ S(i,j) &< \varepsilon,\\ L(i,j) &< \varepsilon, \end{aligned} \tag{2}$$

where $\mathrm{TPV}(i,j)$ denotes the time-space potential between the commuter and the commuter, and the magnitude of the value indicates the likelihood that the commuter will travel in time and space by commuter CB. i, j are commuters. ΔT is time period of study. $S(i,j)$ denotes the difference in distance between commuter i and the place of residence of j. $L(i,j)$ denotes the difference in distance between commuter i and the place of job of j. $t(i,j)$ denotes the time difference between commuters i and j when leaving their place of residence. S is the sets composed by $S(i,j)$. L are the sets composed by $L(i,j)$. T are the sets composed by $t(i,j)/\Delta T$, and ε is the distance threshold, which takes the value of 300–500 m in general. δ is the time threshold. α, β, γ are weighting factors.

According to equation (1), the time-space potential value $\mathrm{TPV}(i,j)$ of commuter CB between commuters i and j is inversely proportional to $S(i,j)$, $L(i,j)$, and $t(i,j)$. Therefore, the smaller the value of $\mathrm{TPV}(i,j)$, the greater the potential for commuters between i and j to take commuter CB travel together. Passengers are similar in space and time of travel, showing a more similar time space of commuting travel. The likelihood that they will share commuter CB travel is higher.

3.2. Solution of Time-Space Potential Model. Firstly, the study area is gridded and the boundaries of the study area are adjusted to generate 5729 1 km × 1 km grids. Secondly, a time window constraint is established to calculate the time-space potential values between commuters in the grid with each cell grid. Finally, all grids in the study area are iterated to obtain the potential value between any commuters. The steps are as follows.

Step 1: the study area is divided into a unit grid of 1 km × 1 km, denoted by U_c, and the unit grid within the entire study area is defined as a set U, and the commuters located in the unit grid form a set P_{C_k}, where $P_{C_k} \subseteq P$, and P is the set of commuters.

Step 2: establish time window constraint TW_t.

Step 3: iterate over all the grids in the study area in terms of the unit cell grid U_c and calculating the values of $S(i,j)$, $L(i,j)$ and $t(i,j)$ among the commuters in each grid.

Step 4: if $T > \delta$ or $S(i,j), L(i,j) > \varepsilon$, then it indicates that i and j do not have the potential for commuter CB.

Step 5: if $T \leq \delta$ and $S(i,j), L(i,j) \leq \varepsilon$, then calculate the time-space potential values between i and j. The entire algorithm process is iterated through all grids until all the time-space potential values of CB between commuters in the study area are calculated.

Algorithm 1 for calculating the time-space potential values of commuter CB is designed according to the calculation process.

3.3. Service Areas and Potential Travel Demand. This section is the core of the paper. Based on the results of the time-space potential value calculation of CB and referring to the literature [21], the threshold of time-space potential value is 0.5. When the time-space potential value is less than 0.5, the distance difference between commuters' residence, workplace, and time difference from home is the smallest. At that time, the commuters have more potential to travel together and the possibility of using the same transportation mode is higher. The unit grids with time-space potential values less than 0.5 are sorted in descending order by the number of commuters. The top 30% of the sorted grids and the last 30% of the sorted grids are taken as the sample set. It is assumed that the 30% unit grids with the higher number of commuters are the high demand area, so that it is equal to "1."

```
Input $left\_lng, left\_lat, right\_lng, right\_lat$, sample set $D = \{x_1, x_2, x_3, \ldots, x_n\}$
for $i$, $j$ in $grid$ do $(i = 1, 2, 3, \ldots, \mathrm{len}(grid))$
if $t[i]$, $t[j]$ in $TW[i]$ do
    $S(i, j) = 2 \times \mathrm{asin}(\mathrm{sqrt}(a)) \times 6371 \times 1000$
    $L(i, j) = 2 \times \mathrm{asin}(\mathrm{sqrt}(b)) \times 6371 \times 1000$
    $t(i, j) = timestamp1[i] - timestamp2[j]$
    if $S(i, j) < \varepsilon$ and $L(i, j) < \varepsilon$ and $t(i, j) < \delta$ do
        $\mathrm{TPV}[i, j] = a \times S(i, j) + b \times L(i, j) + c \times t(i, j)$
Output TPV
```

Algorithm 1: CBTPV algorithm.

The 30% unit grids with lower number of commuters are the low demand area, so that it is equal to "0". Considering the factors that influence commuters' choice of commuter CB travel as the input parameters of the model, construct a logistic regression grid model. Based on the model results, the commuter CB initial service areas and potential travel demand are obtained.

3.3.1. Logistic Regression Model. Logistic regression model is a classification algorithm of machine learning. The algorithm predicts in a classification way and can calculate the probability of each category, which fits the filtering of the grid in the study area of this paper. Firstly, based on the time-space potential model of commuter CB, we initially selected commuters with time-space potential value less than 0.5 and identified their geographical location in the unit grids. Secondly, we choose the average commuting distance, average commuting time, average income, number of bus stations, number of subway stations, average distance from neighboring bus stations, and average distance from neighboring rail stations of commuters in the grids as the input parameters of the logistic regression model. Finally, a binary logistic regression grid model is constructed to predict the unit grid, and the model is solved by SPSS software. The unit grids of high hotspots are filtered and probability values are obtained to mine the potential population of commuter CB.

(i) Logistic regression model theory: logistic regression is the search for the vector of independent variables $X = (X_1, X_2, \ldots, X_n)$ and the binary response Y [21]. The probability of Y belonging to a particular class is modeled.

$$P(X) = \Pr(Y = 1 \mid X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 +, \ldots, + \beta_p X_p. \tag{3}$$

In fact, logistic regression classification is the process of finding a function, mapping the function values for the 0 to 1 interval, and then classifying the data into two categories. Based on continuous exploration, an ideal "unit-step function" is eventually found, and the function value $P(X)$ is mapped to a 0 or 1 class label according to its positivity or negativity. However, the direct design of the step function value in this way is discontinuous, and it is not possible to perform some relevant derivations, which is not conducive to the optimization calculation later. Thus, the Sigmoid function is chosen as the classification function in the Logistic Regression algorithm, and the function expression is as follows:

$$g(z) = \frac{1}{1 + e^{-z}}. \tag{4}$$

The Sigmoid function is an s-shaped curve, with $g(z)$ taking values in the interval [0, 1]; when $z = 0$, $g(z) = 0.5$, when $z \longrightarrow +\infty$, $g(z)$ tends to 1, and when $z \longrightarrow -\infty$, $g(z)$ tends to 0.

Then we have

$$P(X) = \frac{e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 +, \ldots, + \beta_p X_p}}{1 + e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 +, \ldots, + \beta_p X_p}}. \tag{5}$$

The coefficients of the logistic regression model are usually estimated by the maximum likelihood estimation method.

$$L(\beta) = \prod_{i:y_i} p(X_i) \prod_{x':y_i'} p(X_{x'}), \tag{6}$$

where

$$\begin{aligned} \beta &= (\beta_0, \beta_1, \beta_2, \ldots, \beta_p), \\ X_i &= (X_{i1}, X_{i2}, \ldots, X_{ip}). \end{aligned} \tag{7}$$

(ii) Characteristic values: based on the existing basic data, the study is carried out to fully explore the travel demand and service areas of CB. We choose seven important factors as input parameters for the Logistic Regression model, which are strongly influencing commuters to take commuter CB travel.

① Average commuting distance: based on the longitude and latitude information of mobile phone signaling data, we calculate the difference between the Euclidean distance of commuters leaving their place of residence and arriving at their place of job.

② Average commuting time: based on the time difference between the user's departure from the place of residence and arrival at the place of work recorded by the mobile phone signaling data, we consider personal business trips or out of work, etc., and take the average commuting time of three working days in a week as the average commuting time. Then, counting the number of commuters in each unit grid, we calculate the average commuting time of each unit grid.

③ Secondary house prices: considering that the prices of secondary houses can characterize people's income to some extent, based on this, secondary house prices are used as a substitute variable for people's income. The mean value of the price of second-hand houses nearby where commuters reside is calculated as a characteristic to represent the income of commuters.

④ Number of bus stops: invoke Gaode map API interface, use the Python programming language to crawl the latitude and longitude of bus stops in the study areas, and count the number of bus stops in the unit grids.

⑤ Number of rail stops: similar to ④, the Gaode map API interface is retrieved and the Python programming language is used to crawl the latitude and longitude of rail stations in the study area and count the number of rail stations in the unit grids.

⑥ Distance of commuters' neighboring bus stops: the distance of commuters from bus stops and rail stops will influence whether they choose to take CB for commuting. The average value of the shortest distance between bus stops and rail stops in the grid of commuters' neighboring cells is considered as the input parameter of the logistic regression model.

⑦ Distance of commuters' neighboring rail stations: the distance of commuters from the rail station platform will influence whether they choose to take CB for commuting. The average value of the shortest distance of rail stations in the grids of commuters' neighboring units is considered as the input parameter of the logistic regression model.

3.3.2. Service Areas and Potential Travel Demands. Based on the Logistic Regression model, the parameters of the model are input to predict the grids in the study area. Through the theory of the Logistic Regression model, it is known that when $P \geq 0.5$, the prediction result has good predictive value, and the grids are considered as high hotspots grids; on the contrary, when $P < 0.5$, the unit grids are low hotspots grids. Thus, the high hotspots grid area can be used as the commuter CB service areas. And, the commuters that exist in the high hotspot grids are considered as the potential commuter CB travel demand people.

4. Case Study

4.1. Background of the Case. In this study, the commuter CB travel demand and service areas identification method is proposed in the paper. The method is applied to a real case in the central city of Chongqing, China. The distribution of commuters' occupational and residential locations is identified and visualized based on the commuter OD identification algorithm. In Figure 1, it can be seen that commuters' residence is mainly concentrated in the central area of the central city, and the areas are also the commuters' work gathering area.

4.2. Case Results

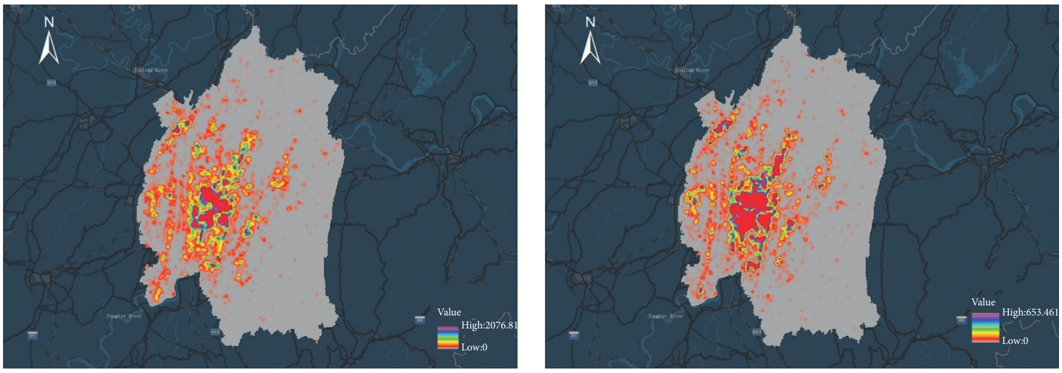
4.2.1. Analysis of the Results of Calculating the Time-Space Potential Value of Commuter CB. Algorithm 1 is designed in Python to calculate the potential values between commuters in the unit grids between 7:00 a.m. and 9:00 a.m. The results are shown in Figure 2. The average value of potential values between commuters in the unit grids is statistically analyzed. And the grids with potential values less than 0.5 in the unit grids are chosen to prepare for the logistic regression model to be established below.

4.2.2. Analysis of Logistic Regression Model Prediction Results. Based on the calculation results of the commuter CB travel potential model, the unit grids with an average travel potential value less than 0.5 (471 units) are chosen and sorted in descending order by the number of commuters in the unit grids. The upper 30% and the lower 30% of the sorted units are taken as the sample set. Since the number of commuters in the upper 30% of the unit grids is higher, they are identified as $Y = 1$, and similarly, the lower 30% of the unit grids are identified as $Y = 0$. The total number of unit grids is 282.

The binary logistic regression model is solved by SPSS software. The fitted results show that the average commuting time, the average distance of neighboring bus stations, the number of bus stations, and the income level had positive effects on the identification of the areas served by commuter CB. The summary table of parameters of the model is shown in Table 1, and the table of prediction accuracy is shown in Table 2.

From Table 1, Wald is 84.817, $P \leq 0.01$. According to the logistic regression theory, it is known that it passed the significance level test and the model is statistically significant. While Cox–Snell R Square is 0.260 and Nagelkerke R Square is 0.346, the fit of the model is high and the model explains the original data at a desirable level.

As can be seen from Table 2, the Sigmoid function takes values in the range of 0-1 interval, with 0.5 as the dividing line. The prediction cannot be used as a commuter CB unit grid in the prediction accuracy rate of 71.6%, the prediction as the service areas has 100 unit grids, and the prediction



(a) (b)

Figure 1: Heat map of where commuters reside and where they work. (a) Heat map of population distribution in the place of residence. (b) Heat map of the population distribution of the workplace.

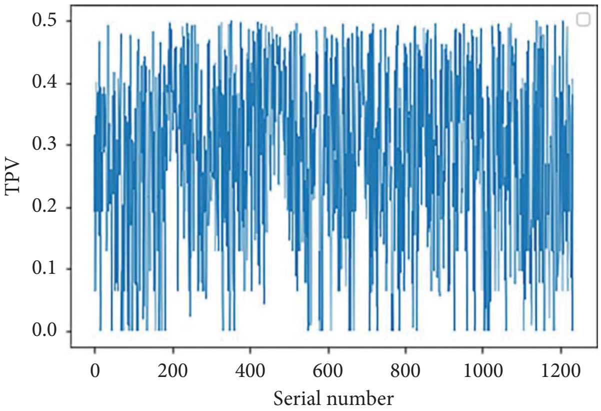


Figure 2: Distribution of time-space potential values of CB.

Table 1: Summary table of model parameters.

Parameters	Wald	*P* value	Cox–Snell *R* square	Nagelkerke *R* square
Values	84.817	≤0.01	0.260	0.346

Table 2: Prediction accuracy.

Actual prediction		Prediction *Y* 0	 1	Accuracy rate (%)
Y	0	101	40	71.6
	1	41	100	70.9
Total accuracy rate (%)				71.3

correct rate is 70.9%. The total prediction accuracy rate is 71.3%, the accuracy rate is 71.43%, the recall rate is 70.92%, and AUC value is 0.811 (as shown in Figure 3). These indicators show that the prediction model is more ideal and the prediction effect is perfect.

Based on the learned model, logistic regression is applied to predict 5729 grids in the central city of Chongqing, China. The machine learning model is solved by SPSS software, and the prediction results are shown in Figure 4.

4.2.3. High Hotspot Grids and Potential Travel Demand. Based on the above analysis of the model results, it can be learned that the prediction results for the area of high

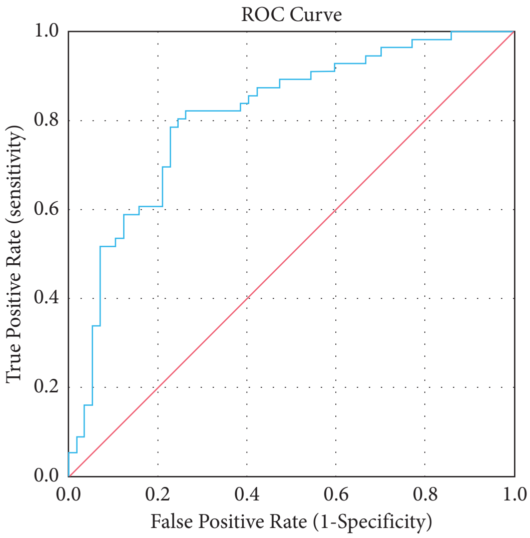


Figure 3: ROC curve.

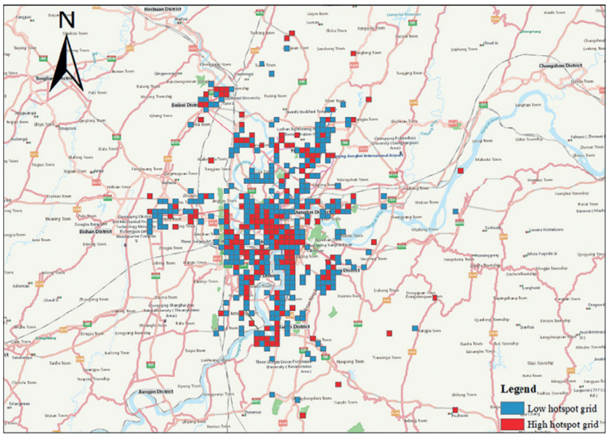


Figure 4: Model prediction results.

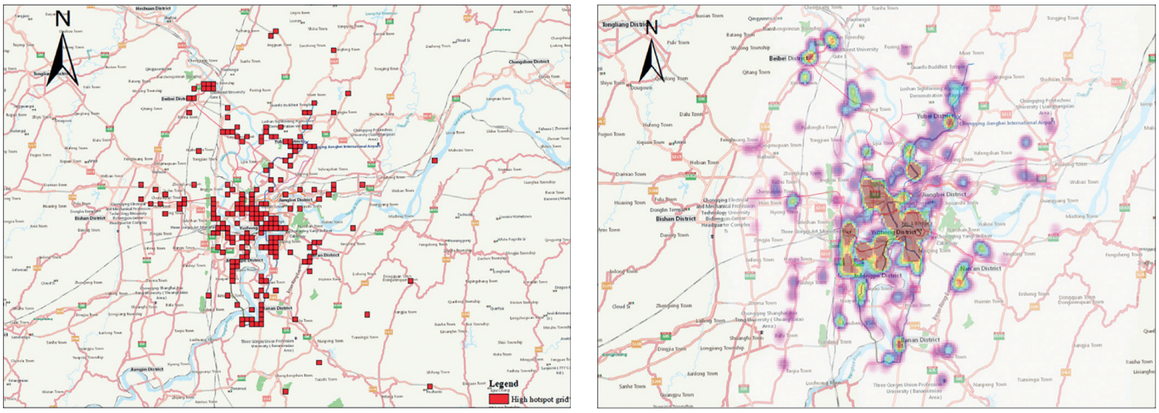
hotspot unit grids (as shown in Figure 5(a)) have advantages for the operation of commuter CB routes. The high hotspot grids areas are considered as the service areas of commuter CB. And, the commuters in the high hotspot unit grids are the potential commuter CB travel demand crowd (as shown in Figure 5(b)).

4.2.4. Examples of Commuter CB Line Planning. By analyzing the distribution of high hotspot grids and travel demand, we randomly chose one high hotspot unit grid each in Shapingba District, Beibei District, and Yubei District of Chongqing, China, as an example to plan commuter CB routes. The commuters in the high hotspot unit grids are considered as potential commuter CB travel demand. The lines information is shown in Table 3.

In this paper, the place of residence is considered as the pickup area and the place of work as the drop-off area. Three randomly selected residential grid areas are surveyed by random sampling to verify the accuracy of the model prediction results. And, in the chosen areas, conduct a questionnaire survey of the commuter CB SP for passengers. The purpose of the SP questionnaire is that the general travel intentions of people in the unit grid represent the travel intentions of potential commuters of CB travel in the unit grid.

One hundred questionnaires are distributed to each of the three chosen areas, for a total of 300 questionnaires, including 95 valid questionnaires for grid ID 4309, 98 valid questionnaires for grid ID 4342, and 94 valid questionnaires for residential grid ID 2654, for a total of 287 valid questionnaires. The results of the questionnaire survey show that the number of passengers in each grid who are inclined to choose commuter CB travel is greater than the predicted number of potential commuter CB travel demand people obtained from the model, which verifies the validity of the model prediction results.

Based on the number and distribution of commuter CB travel demands, the k-means clustering algorithm is used to spatially cluster the travel demand. Since the k-value has a



(a) (b)

Figure 5: High hotspot grids and potential travel demand distribution. (a) High hotspot grids. (b) Potential travel demand distribution heat map.

Table 3: Line information of example.

Line ID	Grid ID of residence	Grid ID of workplace	Distance (km)	Demand (person)
1	4309	3708	10.7	14
2	4342	3900	12	11
3	2654	2597	13	13

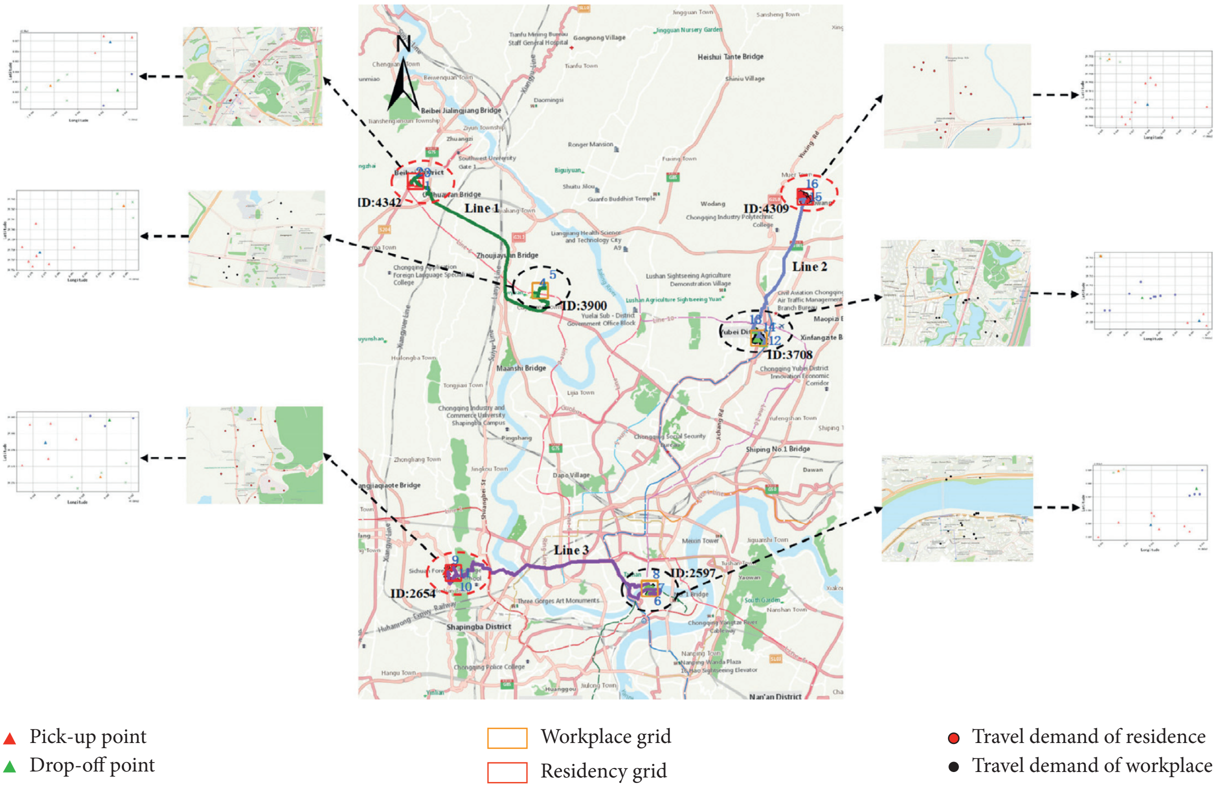


Figure 6: Example of line planning.

large impact on the result of the k-means clustering algorithm, the appropriate k-value is initially determined by applying the Silhouette Coefficient. Then, spatial clustering is carried out, respectively, for residential and workplace travel demand, and line planning is performed for the area based on the clustering results. Through line planning, three vehicles are allocated to meet the passenger travel demand. From the perspective of enterprise operation, the company's constant cost is 240 RMB, the variable cost is 25.23 RMB, and the enterprise's fare revenue is 304 RMB. Without considering the government subsidy, the total revenue is 38.77 RMB, which ensures that the operating enterprise is in a profitable state. The line planning results are shown in Figure 6.

5. Conclusion

Based on the current status of research by many scholars, this paper focuses on the current shortcomings and carries out an in-depth study on the issue of commuter CB travel demand and service areas. The main research contents of this paper are as follows:

(i) Firstly, based on the preprocessing of mobile phone signaling data and commuter OD identification, a commuter CB travel time-space potential model is proposed. Then, the study area is gridded, by designing an algorithm to solve the model.

(ii) Considering commuters who meet certain conditions, Logistic Regression model is applied to analyze the unit grid as the basic cell. We choose the objective factors that influence passengers' choice to take commuter CB as the output parameters of the model and deeply mine the potential population of commuter CB travel demand. We consider the high hotspot grids output of the model as the commuter CB service areas. Finally, using Chongqing, China, as a study case and three routes as examples, the results show that the operating companies are in a profitable state without government subsidies. The case results prove the effectiveness of the method proposed in this paper in practical applications.

In addition, some issues in this paper need to be further discussed:

(i) The data used in this paper are mobile phone signaling data based on COO cellular cell location technology, and there are certain defects in data accuracy. The article chooses to sort the samples of the upper 30% and the lower 30% of the grids, and other methods are also feasible, such as the upper 20% and the lower 20%.

(ii) The paper is not sufficient to justify the value of some model parameters, and it is expected that the parameters of the model can be further studied later to improve the accuracy of the model.

(iii) The operating company can combine the spatial and temporal distribution characteristics of the potential commuter CB travel demand obtained from this paper to introduce intentional routes to specific areas. This way can provide people with convenient travel services.

Data Availability

The data used to support the results of this study are not available because they contain user privacy.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This paper was supported by the National Natural Science Foundation of China (no. 50908150).

References

[1] T. Liu and A. Ceder, "Analysis of a new public-transport-service concept: customized bus in China," *Transport Policy*, vol. 39, pp. 63–76, 2015.

[2] Y. Lyu, C. Chow, V. C. S. Lee, Y. Li, and J. Zeng, "T2CBS: mining taxi trajectories for customized bus systems," in *Proceedings of the 2016 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pp. 441–446, San Francisco, CA, USA, April 2016.

[3] S. Harms and B. Truffe, *The Emergence of a Nationwide Carsharing Co-operative in Switzerland*, Eidg, Anstalt fur Wasserversorgung und Gewasserschutz, Dübendorf, Switzerland, 1998.

[4] K. Tsubouchi, H. Yamato, and K. Hiekata, "Innovative on-demand bus system in Japan," *IET Intelligent Transport Systems*, vol. 4, no. 4, pp. 270–279, 2010.

[5] F. Qiu, W. Li, and J. Zhang, "A dynamic station strategy to improve the performance of flex-route transit services," *Transportation Research Part C: Emerging Technologies*, vol. 48, pp. 229–240, 2014.

[6] L. Scott, M. Lee-Gosselin, A. Sivakumar, and J. Polak, "A new approach to predict the market and impacts of round-trip and point-to-point carsharing systems: case study of London," *Transportation Research Part D: Transport and Environment*, vol. 32, pp. 218–229, 2014.

[7] K. An and H. K. Lo, "Two-phase stochastic program for transit network design under demand uncertainty," *Transportation Research Part B: Methodological*, vol. 84, pp. 157–181, 2016.

[8] T. Liu, A. Ceder, A. Ceder, R. Bologna, and B. Cabantous, "Commuting by customized bus: a comparative analysis with private car and conventional public transport in two cities," *Journal of Public Transportation*, vol. 19, no. 2, pp. 55–74, 2016.

[9] Y. Lyu, C.-Y. Chow, V. C. S. Lee, J. K. Y. Ng, Y. Li, and J. Zeng, "CB-Planner: a bus line planning framework for customized bus systems," *Transportation Research Part C: Emerging Technologies*, vol. 101, pp. 233–253, 2019.

[10] J. Zhong, H. Bai, and Z. Wu, "Division method of urban customized bus passenger flow distribution area for commuter- oriented demand," in *Proceedings of the 2018 World Transport Conference*, Geneva, Switzerland, August 2018.

[11] Y. Cheng, A. Huang, G. Qi, and B. Zhang, "Mining customized bus demand spots based on smart card data: a case study of the beijing public transit system," *IEEE Access*, vol. 7, pp. 181626–181647, 2019.

[12] Q. Yu, W. Li, and H. Zhang, "Mobile phone data in urban customized bus: a network-based hierarchical location selection method with an application to system layout design in the urban agglomeration," *Sustainability*, vol. 12, 2020.

[13] Q. Liu, Q. Li, and C. Tang, "A visual analytics approach to scheduling customized shuttle buses via perceiving passengers' travel demands," 2020, https://arxiv.org/abs/2009.02538.

[14] J. Zhang, D. Z. Wang, and M. Meng, "Analyzing customized bus service ona multimodal travel corridor: an analytical

modeling approach," *Journal of Transportation Engineering*, vol. 143, no. 11, 2017.

[15] L. Tong, L. Zhou, J. Liu, and X. Zhou, "Customized bus service design for jointly optimizing passenger-to-vehicle assignment and vehicle routing," *Transportation Research Part C: Emerging Technologies*, vol. 85, pp. 451–475, 2017.

[16] W. Huang, W. Jin, J. Huang, and B. Han, "Pricing problem of customized bus under different market strategies," *Journal of Guangxi Normal University*, vol. 2, no. 2, 2018.

[17] J. Ma, Y. Yang, and W. Guan, "Large scale demand driven design of a customized bus network: a methodological framework and beijing case study," *Journal of Advanced Transportation*, vol. 2017, Article ID 3865701, 14 pages, 2017.

[18] D. Li, X. Ye, and J. Ma, "Empirical analysis of factors influencing potential demand of customized buses in Shanghai, China," *Journal of Urban Planning and Development*, vol. 145, no. 2, 2019.

[19] X. Y. Tang, T. Zhou, B. C. Lu, and Z. G. Gao, "A commuting OD matrix training method based on mobile phone data," *Transportation System Engineering and Information*, vol. 16, no. 5, pp. 64–70, 2016, in Chinese.

[20] E. H. Geng, *Analysis of Ride-Sharing Potential Based on Bus Data*, University of Electronic Science and Technology, Chengdu, China, 2017, in Chinese.

[21] C. Yu, C. Xu, X. Ding, and L. Zeng, "Optimizing location of car-sharing stations based on potential travel demand and present operation characteristics: the case of chengdu," *Journal of Advanced Transportation*, vol. 2019, Article ID 7546303, 13 pages, 2019.

Hindawi
Journal of Advanced Transportation
Volume 2021, Article ID 9922660, 13 pages
https://doi.org/10.1155/2021/9922660

Research Article

Forecasting Passenger Flow Distribution on Holidays for Urban Rail Transit Based on Destination Choice Behavior Analysis

Enjian Yao,[1] **Junyi Hong**,[1] **Long Pan**,[2] **Binbin Li**,[3] **Yang Yang**,[1] **and Dongbo Guo**[1]

[1]*School of Traffic and Transportation, Beijing Jiaotong University, Beijing 100044, China*
[2]*College of Metropolitan Transportation, Beijing University of Technology, Beijing 100124, China*
[3]*School of Traffic and Transportation, Chongqing Jiaotong University, Chongqing 400074, China*

Correspondence should be addressed to Enjian Yao; enjyao@bjtu.edu.cn

Received 9 March 2021; Accepted 31 August 2021; Published 24 September 2021

Academic Editor: Steven Chien

Passenger travel flows of urban rail transit during holidays usually show distinct characteristics different from normal days. To ensure efficient operation management, it is essential to accurately predict the distribution of holiday passenger flow. Based on Automatic Fare Collection (AFC) data, this paper explores the passengers' destination choice differences between normal days and holidays, as well as one-way tickets and public transportation cards, which provides support for variable selection in modeling. Then, a forecasting model of holiday travel distribution is proposed, in which the destination choice model is established for representing local and nonlocal passengers. Meanwhile, explanatory variables such as land matching degree, scenic spot dummy, and level of service variables are introduced to deal with the particularity of holiday passengers' travel behavior. The parameters calibrated by the improved weighted exogenous sampling maximum likelihood (WESML) method are applied to predict passenger flow distribution in different holiday cases with annual changes in the metro network, using the data collected from Guangzhou Metro, China. The results show that the proposed model is valid and performs better than the other comparable models in terms of forecasting accuracy. The proposed model has the capability to provide a more universal and accurate passenger flow distribution prediction method for urban rail transit in different holiday scenarios with network changes.

1. Introduction

With the development of the economic level, the travel activities and frequencies of urban residents continue to increase, which leads to the rapid growth of urban residents' demand for urban public transport. Urban rail transit has developed rapidly in recent years, and its superiority of traffic volume, speed, and punctuality are popular among people, which helps spur a boom in urban rail construction [1]. In recent years, a large number of new lines have opened and connected to the metro network, making the network operation effect of many cities particularly evident, significantly affecting regional accessibility and passenger flow distribution in the metro network. Furthermore, in regard to holidays, because of the exceptional flexibility of departure time and the diversity of destinations, the passenger travel characteristics are quite distinct from normal days, and the spatiotemporal distribution of holiday travel demand presents complex characteristics [2, 3].

With the rapid change of the metro network, the operations have undergone quantitative and qualitative changes [4]. The particularity of holidays also aggravates travel demand's complexity, which poses a significant challenge to the metro system. Besides, the same holiday only occurs once a year, which is not conducive to study the characteristics in terms of lacking data sources. Therefore, to effectively organize the large passenger flow and alleviate traffic congestion during holidays, it is essential to accurately predict the distribution of holiday passenger flow, which is the basis of a reasonable train operation plan-making and the development of passenger flow induction strategy.

The traditional four-step methods and their modification models have been widely used in passenger flow distribution forecasting. It mainly includes the aggregate model method based on statistical rules of historical data and the disaggregate model method based on behavior analysis.

In the study of the aggregate model methods, many researchers have investigated the gravity model by improving it in different contexts. Grosche et al. [5] proposed two gravity models to estimate the air passenger flow between city-pairs. They introduced geoeconomic variables describing the general economic activity and geographical characteristics as independent factors. Wang et al. [6] combined the gravity model considering the distance and free-flow travel time with the Fratar method to predict the seed O-D matrix of the expressway. More recently, Ren et al. [7] proposed three types of land-use function complementarity indices introduced into spatial interaction to improve the gravity model. In these studies, the appropriate variables are introduced to modify the model. Besides, the constrained gravity model is also used as a researching point. Tsekeris and Stathopoulos [8] used a doubly constrained gravity model that additionally incorporates the intraperiod evolution for forecasting the dynamic trip distribution. Jin et al. [9] proposed an O-D estimation model based on the doubly constrained gravity model, where the comparison of singly and doubly constrained models was made. However, the aggregate gravity model tends to overestimate when the distance-deterrence function is small, and the variables are usually less and simple, which cannot reflect the forming mechanism of passenger flow and travel behavior objectively.

The disaggregate model can reveal the internal mechanism of the passengers' destination choice from the perspective of behavior interpretation by establishing definable variables. Specifically, previous studies on the disaggregate model have focused on travel behavior analysis and demand forecasting. For example, in the research of influencing factors of travel behavior, Tsirimpa et al. [10] proposed a multinomial logit model and a mixed multinomial logit model to examine the impact of information acquisition on switching travel behavior. Yang et al. [11] proposed multinomial and nested logit models to analyze battery electric vehicle drivers' charging and route choice behaviors. Nguyen-Phuoc et al. [12] adopted a multinomial logit model to explore factors affecting changes in the event of major public transport disruptions. In addition, the discrete choice modeling technology based on random utility-based is mainly used for destination choice modeling. Faghih-Imani [13] used a multinomial logit model to study the decision process of identifying destination locations at a bicycle station. Kelly [14] built multinomial logit models to analyze the destination choice behaviors of pedestrians within an entire region. Orvin [15] developed a random parameter latent segmentation-based logit model to investigate trip destination choice behavior of the dockless bike-sharing users. These studies show that individual attributes and alternative factors influence passenger behavior and the decision process, assisting transit agencies in getting management guidance.

Focusing on the demand forecasting, Timmermans [16] proposed a model combining transportation mode selection and destination selection and predicted shopping-oriented travel. To strengthen the forecasting power, Jovicic and Hansen [17] constructed a nested logit model, where logsums integrate generation, distribution, and mode choice models as submodels. Ashiabor et al. [18] developed the nested and mixed logit model to estimate county-to-county travel demand. Travel time, cost, and traveler's household income were used in the explanatory variables. Furthermore, recent studies [19] proposed a multistage demand forecasting model that considers the discrete choice approach, such as the binomial and multinomial logit model, for each decisional level. Moreover, Li [20] presented a new itinerary-based nonlinear demand estimator that estimates the distribution of demand based on a nested logit model. These studies contribute to the accurate prediction of travel demand with the improved disaggregate model. However, it is usually necessary to use questionnaires, such as the stated and revealed preference surveys, to obtain the data that include individual and alternative attributes for studying the behavioral characteristics. When applied to prediction, it is easy to be restricted by data conditions and difficult to use effectively.

In addition, many emerging data mining technologies and methods are used to study traffic or passenger flow demand. Ye and Wen [21] proposed a destination choice model based on link flows by constructing algorithms observing the detected data from part of the links. By using data mining, Wang et al. [22] developed cell phone location tracking algorithms to track cross-region traffic activities and derived the O-D traffic flow and travel demand. In the machine learning approaches, Wang et al. [23] designed a grid embedding network via graph convolution and established a multitask learning network for forecasting the demands of O-D pairs in ride-hailing. Although the prediction accuracy of the data-driven approach depending on long-term collection may be higher, it is hard to apply the network structure changes because of lacking the newly added stations' data in the metro. Moreover, it is often a black-box process that does not illustrate the internal behavior mechanism.

Generally, due to the holidays that occurred only once a year, it is not easy to continuously collect stable and long-term data. And with the rapid development of the metro system, the network structure of holidays usually changes every year, which makes it hard to use the statistical models for prediction. Previous relevant studies focused on researching the passenger flow on normal days. However, little work has explored passengers' choice behavior to construct special variables to effectively forecast the scenarios of holidays in the metro system. Besides, since the source of disaggregate data limits the forecasting application, new data sources are considered to replace the conventional questionnaires in this paper.

At present, the Automatic Fare Collection (AFC) system is widely adopted in the urban rail transit system, which is the main support data in this paper. Under the premise of ensuring validity, this paper applies the aggregate data obtained by the AFC to the disaggregate model by modifying the maximum likelihood estimation method, which overcomes the difficulty of getting the disaggregate model data.

Based on fully exploiting the holiday passenger travel rules and considering the differences in the choice behavior of different ticket type passengers, this paper constructs a holiday passenger flow distribution prediction model, in which some novel explanatory variables (such as land matching degree) are introduced. The proposed model structure can not only be suitable for the changes of urban rail transit network structure but also take into account the unique characteristics of holidays so as to have good interpretability.

The remainder of this paper is organized as follows. In the next section, the holidays' data collection effort and passenger flow characteristics are analyzed herein. Then, the modeling methodology and the explanatory variables of the utility function are described. After that, the proposed model is estimated and applied to the holiday distribution forecasting with comparisons of other traditional methods. Finally, concluding remarks are presented in the last section.

2. Data and Passenger Flow's Characteristic

2.1. Data. Urban rail transit adopts the AFC system to implement management methods such as ticketing, ticket checking, and billing. The data are gathered and transmitted into the center and automatically store passenger travel information. The data types are shown in Table 1. Under such limited data conditions and types, how to use them to construct a forecasting model suitable for the holiday scenario is the primary goal. In data processing, the data cleaning has been done by identifying outliers, such as judging whether the enter and exit stations are inconsistent, whether the enter and exit time, and the in-train time are reasonable. Besides, the stations are regarded as transportation analysis zones (TAZs) in the urban rail transit system. The boarding (origin) and alighting (destination) stations of passengers' trips can be obtained from the AFC system directly.

There are eight lines and 140 stations in Guangzhou Metro by the beginning of 2016. The daily average of raw data amounts to more than 4 million that need further processing. And more than one million passengers use one-way tickets per day during New Year's Day, which is almost 1.84 times the weekdays. Compared with January 1, 2016, there are seventeen new stations and three new lines connected to the network on January 1, 2017. The road network structure has tremendous changes.

2.2. Passenger Flow's Characteristics. Based on Guangzhou Metro's AFC data, the passenger flow of each station during the New Year's Day holiday from 2016 to 2017 is collected, and some travel characteristics have been found. The passenger flow, for instance, is closely related to the nature of land-use and the intensity of development around stations.

As shown in Figure 1, the entrance passenger flows of four typical stations from December 30, 2015, to January 4, 2016, are given. The passenger flow of Zhujiangxincheng station, which is dominated by office areas, declined significantly during the New Year's Day. Similarly, the passenger flow of Dashadong station also decreased, as residential areas surround there. However, Guangzhouta and Beijinglu stations' passenger flow increased significantly during the holidays, with the main areas, respectively, surrounded by scenic spots and commercial districts.

Similarly, the passenger flow, from the origin station to the destination station (O-D station) during the holidays, shows different distinct characteristics, compared with weekdays. As shown in Figure 2, there are different passenger flow trends between O-D stations with different land-use types, and some of which increased significantly in holidays, while others, such as residential stations to office stations, dropped significantly.

From another perspective, there are also great differences in the distribution of people who use one-way tickets and public transportation cards during holidays. Generally, many one-way passengers are nonlocal passengers, who tend to go to scenic spots, business districts, and hub stations. In contrast, transportation card passengers are mostly local residents, whose travel purposes are diversified. This characteristic of choice behavior is especially evident during holidays. As shown in Figure 3, the passenger flow of one-way tickets and public transportation cards at Guangzhouta and Beijinglu stations has increased, while the growth rate of one-way ticket is significantly higher, indicating that the stronger attraction of one-way ticket passengers.

Furthermore, other characteristics can also be obtained by analyzing the passenger flow. For example, the O-D passenger flow on the same line is usually larger than that on different lines. And in the case of satisfying the purpose of passengers, they would give priority to the destination with a short ride time and transfer time. However, these features are influenced by many factors. They should be reflected in some explanatory variables to analyze how various factors jointly affect the behavior and improve the subsequent forecasting performance when modeling. Next, the approach considering passenger flow's characteristics of holidays is introduced in detail.

3. Methodology

Considering that the metro network scale is rapidly developing, the spatial passenger flow distribution of O-D stations also changes fast. New stations divert the passenger flow of old stations, and it is not easy to obtain the development data of all O-D pairs in time series, especially for newly added stations. Therefore, based on the above analysis of passenger flow's features, this paper constructs a destination choice model to describe the characteristics of passengers. Then, a forecasting model of holiday passenger flow distribution is developed, which is suitable for the structural change of the network and does not depend on long-term data collection. Meanwhile, considering different passengers' characteristics, the utility functions for passengers who use one-way tickets and public transportation cards are constructed separately.

3.1. Model Structure. The theory of random utility maximization refers to the alternatives in which traffic behavior decision-makers choose the most effective ones in their

Table 1: The Automatic Fare Collection data types.

Name	Field type	Form
Card type	Varchar	One-way ticket, public transportation card, etc.
Card number	Int	1000139^{***}
Enter line	Varchar	Line 1, Line 2, etc.
Enter station name	Varchar	Guangzhou East, Donghu station, etc.
Enter time	Datetime	"2016-01-01 08 : 00 : 00"
Exit line name	Varchar	Line 1, Line 2, etc.
Exit station name	Varchar	Guangzhou East, Donghu station, etc.
Exit time	Datetime	"2016-01-01 08 : 00 : 00"

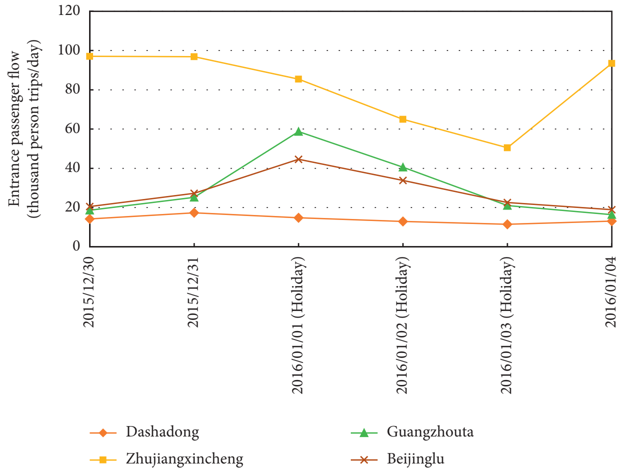


Figure 1: Entrance passenger flow of typical stations during New Year's Day.

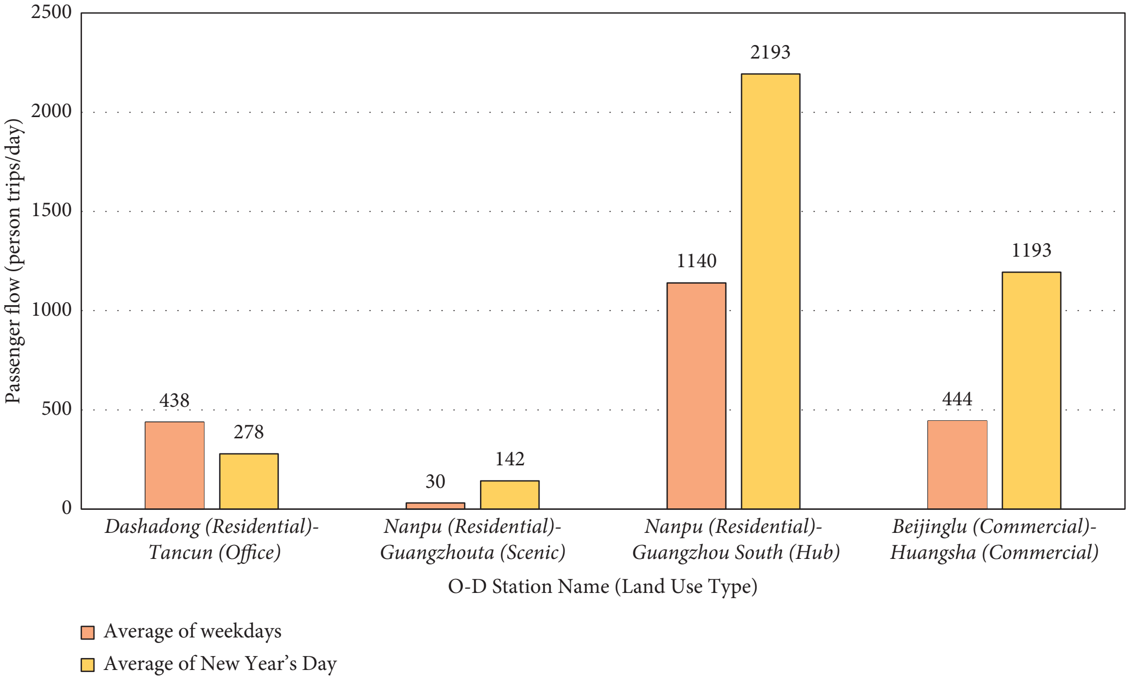
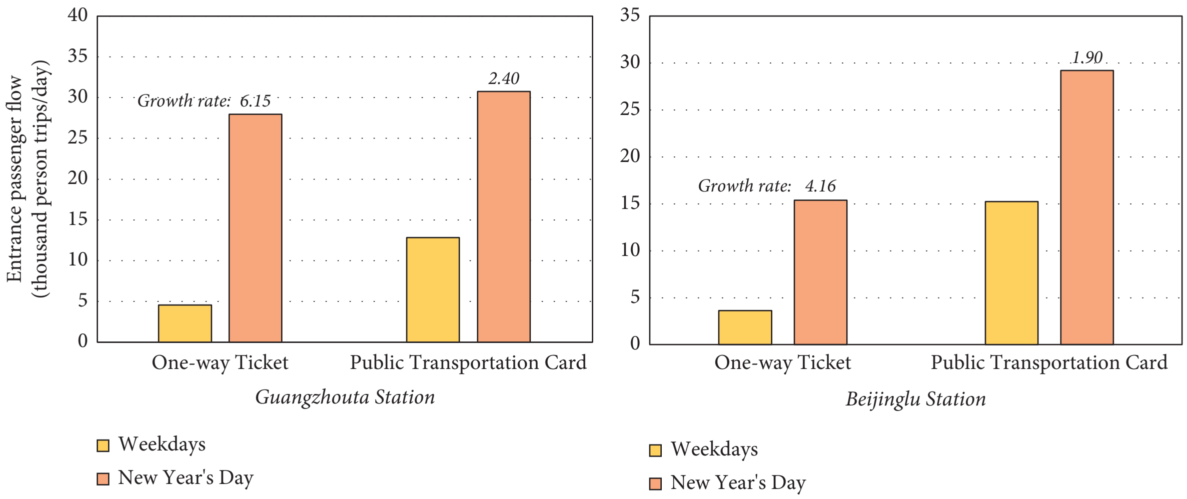


Figure 2: Comparison of O-D passenger flow during weekdays and New Year's Day in 2016.

choice sets under certain conditions. If the destination choice sets of passengers from station i are A_i and the utility of the alternative n is U_{in}, the requirement that the passengers select the destination j from A_i is $U_{ij} > U_{in}$. Among them, the utility function U has divided into two parts: a deterministic term V_{ij} and an error term ε_{ij}.



(a)

(b)

Figure 3: Comparison of passenger flow between one-way ticket and public transportation card in 2016. (a) Guangzhouta station. (b) Beijinglu station.

Therefore, the utility function U_{ij} can be formulated as follows:

$$U_{ij} = V_{ij} + \varepsilon_{ij} = \sum \beta^k X_{ij}^k + \varepsilon_{ij}, \tag{1}$$

where β^k is an estimable parameter of attribute k; X_{ij}^k is an observable attribute as the explanatory variable; ε_{ij} is the error term that is used to address the unobserved factors that influence the choices taken by the passengers.

The researcher observes some attributes of the alternatives as faced by the decision maker, labeled X_{ij}, and can specify a function that relates these observed factors to the decision maker's utility [24]. The term ε_{ij} is treated as random, and it captures the factors that affect utility but are not included in V_{ij}. When the error term ε_{ij} obeys the independent Gumbel distribution, multinomial logit (MNL) models can be derived. For the origin station i, the probability for choosing j is calculated as follows:

$$P_{ij} = \frac{\exp\left(V_{ij}\right)}{\sum_{n\varepsilon A_i} \exp\left(V_{\text{in}}\right)}, \quad j\varepsilon A_i. \tag{2}$$

Equation (2) is the destination choice model. The probability that a passenger chooses another station as the destination can be calculated. The production trips from each station are then distributed to all other stations based on the choice of probability destination. That is, the passenger flow distribution q_{ij} from the origin station i to the destination station j is computed. The formula is shown as follows:

$$q_{ij} = Q_i \cdot P_{ij}, \tag{3}$$

where Q_i is the entrance passenger flow in station i.

Considering the different levels of sensitivity of travel characteristics of different types of passengers in the particularity of holidays, two utility functions in the proposed model are constructed with passengers who use one-way tickets and public transportation cards for representing local and nonlocal passengers. The trip distribution is applied separately for each ticket type of passengers who have characteristic travel behavior, with different model parameters. Then, the distribution results of the two ticket types are added together. The formula is shown as follows:

$$q_{ij} = q_{ij}^{\text{one}} + q_{ij}^{\text{ptc}}, \tag{4}$$

where q_{ij}^{one} is the one-way ticket passengers' distribution prediction; q_{ij}^{ptc} is the public transportation card passengers' distribution prediction.

Equation (4) is the forecasting model of passenger flow distribution. However, it is a singly constrained model so far. There is no guarantee that the sum of the passenger flow from each station to the destination station j is equal to the attracted trips of station j. Therefore, it is necessary to modify the travel flow to enforce constraints between total origins and destinations. The Fratar method is widely used in distribution adjustment due to its fast convergence speed and high calculation accuracy. The idea of the Fratar method is a distribution of horizon year trips from a zone that is proportional to the base year trip distribution pattern modified by the growth factors of the zones under consideration [25, 26]. Therefore, this paper uses the Fratar method for equalization processing. The approach is shown as follows:

$$\begin{aligned} q_{ij}^{m+1} &= q_{ij}^{m} \cdot F_{Oi}^{m} \cdot F_{Dj}^{m} \cdot \left(\frac{L_i + L_j}{2}\right), \\ L_i &= \frac{O_i^m}{\sum_j q_{ij}^m \cdot F_{Dj}^m}, \\ L_j &= \frac{D_j^m}{\sum_i q_{ij}^m \cdot F_{Oi}^m}, \end{aligned} \tag{5}$$

where q_{ij} is the passenger flow of station i to station j; F_{Oi} is the growth rate of the entrance passenger flow in station i; F_{Dj} is the growth rate of the exit passenger flow in station j; L_i is the adjustment coefficient of station i; L_j is the adjustment coefficient of station j; O_i is the entrance passenger flow in station i; D_j is the exit passenger flow in station j; m is the m-th iteration.

3.2. Model Specifications. Although personal characteristics affect destination choice, it is unable to obtain personal attributes data from the AFC directly. Therefore, seven indexes as the utility function of characteristic variables that could be extracted from the urban rail transit network are considered in the destination choice model, including in-vehicle travel time, transfer time, station position relationship, and matching degree of land-use types. The seven variables are mainly used to characterize three categories of explanatory attributes, namely, the accessibility of the destination, the attractiveness of the destination, and matching degree of O-D stations, through which the choice behavior mechanism of passengers can be characterized.

According to the choice behavior characteristics of the one-way ticket and the public transportation card passengers, and through the multiple calibration experience of the model, the utility functions V_{ij}^{one} and V_{ij}^{ptc} of the destination choice model are constructed, as shown in equations (6) and (7), respectively:

$$V_{ij}^{\text{one}} = \beta_1 D_j + \beta_2 Z_{ij} + \frac{\beta_3 T_{ij}^{\text{train}}}{3600} + \frac{\beta_4 N_{ij}^{\text{trans}}}{3600} + \beta_5 G_{ij} + \beta_6 S_{ij} + \beta_7 L_j, \tag{6}$$

$$V_{ij}^{\text{ptc}} = \beta_8 D_j + \beta_9 Z_{ij} + \frac{\beta_{10} T_{ij}^{\text{train}}}{3600} + \frac{\beta_{11} N_{ij}^{\text{trans}}}{3600} + \beta_{12} G_{ij} + \beta_{13} S_{ij}, \tag{7}$$

where $\beta_1 - \beta_{13}$ is the parameter to be calibrated for each variable; D_j is the exit passenger flow of destination station j, ten thousand person trips; Z_{ij} is the matching degree of land-use type; T_{ij}^{train} is the in-vehicle travel time from the origin station i to the destination station j, second; N_{ij}^{trans} is the transfer time from the origin station i to the destination station j, second; G_{ij} is a dummy variable, and if the sum of trip generation at origin station i and the attraction at destination station j is larger than a specific scale, the value is 1; S_{ij} is a dummy variable, and if the origin station i and destination station j are in the same line, the value is 1; L_j is a dummy variable, and if the land-use type of destination station j is scenic, commercial, or hub, the value is 1.

For one thing, these variables are introduced to facilitate data acquisition, and for another, the characteristics of holidays are considered so as to improve the interpretability and prediction effect of the model further. It should be noted that the travel cost is a sensitive variable to influence the choice behavior, which was included in the variable sets at the beginning. However, when the variables are checked for multicollinearity, the travel cost shows a strong correlation with the travel time. Therefore, the travel cost was eliminated in the utility functions. Compared to one-way ticket passengers, the public transportation card utility functions do not have the variable L_j, as adding this variable would reduce the model's accuracy.

Moreover, the acquisition of the matching degree of land-use types Z_{ij} and the scenic destination station variables L_j need to be additionally explained. The distribution of passenger flow between stations is closely related to land-use nature around the station, especially the significant difference between holidays and normal days. It is necessary to quantify the impact of land-use interaction. Therefore, Z_{ij} is constructed to describe the degree of attraction between different types of stations. Based on this, the metro stations need to be clustered to determine the category of the station first.

Due to the land-use properties are a relatively stable indicator and it usually shows a certain relationship with the passenger flow characteristics, the K-means clustering method is used to classify the stations of the whole network of Guangzhou Metro. K-means is a vector quantization method that is popular for cluster analysis in data mining [27]. Through the analysis of passenger flow characteristics, the morning and evening peak flow has a greater correlation with the nature of land-use around stations. And the proportion of one-way tickets and all-day passenger flow at comprehensive transportation hubs is usually larger, while the passenger flow at commercial and scenic stations tends to increase significantly during holidays. Therefore, the five variables are used as inputs for clustering as shown in Table 2. In the clustering research of metro stations, the stations are usually divided into five categories according to weekday travel data [28, 29]. However, since the research scenarios are aimed at holidays, we set eight cluster numbers as preset categories according to the land-use and application requirements of the model. The clustering results are shown in Table 3 (figures in brackets denote the sum number of clustering stations), and they are representative and matched with the preset types.

Therefore, the value of L_j can be obtained directly through the clustering results. Besides, the matching degree of land-use type Z_{ij} needs further processing. Based on the above clustering results, the average O-D passenger flow with different cluster types could be calculated. Then, the logarithm function is used to normalize the values of various types to differentiate passenger flow better. The formula is as follows:

$$Z_{ij} = \frac{\left(\ln Q_{ij} - \ln \min_{i,j} Q_{ij}\right)}{\left(\ln \max_{i,j} Q_{ij} - \ln \min_{i,j} Q_{ij}\right)}, \tag{8}$$

where Z_{ij} is land matching degree from type i to type j; Q_{ij} is the average O-D passenger flow of the stations from type i to type j.

A case result of Z_{ij} is shown in Table 4 (the vertical column indicates the type of the origin station, and the horizontal row indicates the type of the destination station), where the value from Type1 to Type 1 is zero. This means that the passenger flow is the lowest of all type pairs, mainly because the attraction between residential stations is less during holidays in all type

Table 2: The variables for *K*-means.

Variables	Description
Morning peak hour factor	The passenger flow of morning peak hour (7 : 00–9:00) divided by all-day passenger flow
Evening peak hour factor	The passenger flow of evening peak hour (17 : 00–19 : 00) divided by all-day passenger flow
Proportion of one-way ticket	The proportion of passengers using one-way ticket
All-day passenger flow	The all-day passenger flow in the stations
Passenger flow growth rate	The passenger flow of the holiday divided by the weekdays before the holiday

Table 3: Station types' clustering results based on the surrounding area's land-use.

Type	Attribute	Description	Clustering results
Type 1	Residential	The station is surrounded by residential areas	Nanpu, Sanxi, Dongpu, etc. (30)
Type 2	The majority of residential area	The station is surrounded by majority of residential areas	Shiqiao, Meihuay, Xicun, etc. (25)
Type 3	Office	The station is surrounded by office areas	Haizhuguagnc, Quzhuang, etc. (20)
Type 4	The majority of office area	The station is surrounded by majority of office areas	Donghu, Ximenkou, etc. (21)
Type 5	Comprehensive	The station is surrounded by various types of land-use	Shibi, Fangcun, Shiergong, etc. (28)
Type 6	Commercial	The station is surrounded by commercial areas	Jinzhou, Jinronggaoxinqu, etc. (14)
Type 7	Scenic	The station is close to scenic spot	Diyong, Guangzhouta, etc. (14)
Type 8	Hub center	The station functions as transportation hub center	Guagnzhou South, Airport South, etc. (5)

Table 4: Land matching degree values of different types of O-D stations.

	Type 1	Type 2	Type 3	Type 4	Type 5	Type 6	Type 7	Type 8
Type 1	0.00	0.08	0.24	0.08	0.10	0.18	0.28	0.55
Type 2	0.09	0.15	0.28	0.15	0.16	0.24	0.29	0.63
Type 3	0.09	0.17	0.35	0.18	0.18	0.20	0.27	0.61
Type 4	0.22	0.25	0.28	0.29	0.22	0.22	0.29	0.64
Type 5	0.11	0.15	0.24	0.16	0.22	0.22	0.36	0.60
Type 6	0.17	0.23	0.25	0.18	0.23	0.14	0.26	0.59
Type 7	0.28	0.26	0.25	0.19	0.34	0.22	0.39	0.59
Type 8	0.51	0.59	0.61	0.55	0.57	0.56	0.58	1.00

pairs. In contrast, the connections between transportation hubs are strengthened, reflected in the maximum value from Type 8 to Type 8.

3.3. Parameter Estimation. For the parameters β in equation (1), the personal travel survey is generally performed by the simple random sampling method to obtain the disaggregate type data of the individual choice, thereby using the maximum likelihood estimation method to calibrate the parameters. However, in this paper, the aggregate data obtained by the AFC should be transformed into the disaggregate form for application in the destination choice model. When being applied, it needs methods to deal with the original aggregate data. Yao and Takayuki [30] proposed an integrated model that combines estimation across multiple data sources such as SP, RP, and aggregate data. Therefore, the maximum likelihood estimation method is improved by introducing a weight factor to realize the application of AFC data in the destination choice model' calibration.

Manski and Lerman [31] proposed a weighted exogenous sampling maximum likelihood (WESML) method, introducing weights into log-likelihood functions to calibrate the bias between the sample and population data. It can be expressed as follows:

$$L(\theta) = \sum_{i\varepsilon A_n}\sum_{n}\delta_{\text{in}} w_i \ln\left(P_{\text{in}}\right), \tag{9}$$

$$w_i = \frac{Q_i}{H_i}, \quad i\varepsilon C, \tag{10}$$

where δ_{in} is 1 if the passenger n chooses selected branch i as destination and 0 otherwise; w_i is weights; Q_i is the proportion of the selected branch i in the population; H_i is the proportion of i in the sample.

To improve the practicability of the method, Cosslett's research [32] proves that it can be transformed as follows:

$$w_i = \frac{Q_i}{N_i/N}, \tag{11}$$

where N_i is the data amount of the selected branch i; N is the sum of the data amounts of the respective selected branch.

However, in terms of urban rail transit, passengers with the same origin and destination station have the same characteristics; that is, they all make the same choice for the destination. Thus, the amount of O-D passenger flow can be expressed as the selection result of individuals. The weight factor is suited for adjusting the likelihood function of the dataset. Therefore, according to the characteristics of the data that can be extracted, equation (11) is corrected as follows:

$$w_i = \frac{q_i}{\sum q_i} \cdot R, \tag{12}$$

where q_i is the O-D passenger flow in the selected branch i; R is the number of individuals, that is, the sum of O-D station pairs.

4. Results and Analysis

4.1. Model Estimation and Analysis. In the construction of the selection set, there are 140 stations in the 2016 New Year's Day. That is to say, 139 stations should be put into the alternative set except for the real choice of each traveler. However, for general disaggregate models, the size of the alternatives is too large, which would affect the speed of model estimation and is not conducive to application. Ben-Akiva and Lerman [33] demonstrated that the consistency of model parameters is not lost when extracting subselective branches for parameter estimation in the selection set. Therefore, this paper constructed the subselection set by randomly extracting nine stations from the alternative set. It could reduce the difficulty of calibration and increase the operability while ensuring the consistency of the model's calibrated parameters.

In the process of parameter calibration, the values of seven variables are obtained in combination with the network topology and train operation plan of Guangzhou Urban Rail Transit. By using the parameter estimation method described in the section before, the undetermined parameters of the utility function are calibrated. Especially, after several tests, the scale dummy variable G_{ij} was set to 1 if it exceeds 7,000 person trips. The calibration results of the New Year's Day are shown in Table 5 as a study case. All absolute t-values are greater than 1.96, indicating statistical significance and variables' validation. Moreover, the adjusted ρ^2 of this model is over 0.2, which can be regarded as a satisfactory goodness-of-fit [34].

The estimated parameters are provided with practical significance and expected signs in the sense of explaining passenger destination choice behavior in either the one-way ticket model or public transportation card model. An obvious example is that the parameter of destination attraction variable is positive, indicating that the greater attraction of destination station is, the more passengers choose.

As for the negative parameters of travel time and transfer time, the longer the travel time and transfer times are, the less probability of destination station would be chosen, which is consistent with common sense and inversely proportional to destination choice preference. Moreover, the units are the same, but the estimated parameters are not close, which means the travelers have different perceptions. The trade-off between travel and transfer time shows that an increase of 10 minutes in transfer time is equivalent to an increase of 68 minutes in travel time for one-way ticket passengers and 55 minutes for public transportation card passengers in the case of New Year's Day. It reveals that travelers have a significant negative impact on lengthy transfer times. For public transportation card passengers, the absolute parameters of travel time and transfer time are both larger than the one-way ticket passengers, indicating that the passengers who used the card care more about the time when other variables remain unchanged.

Besides, the land matching degree's parameter is positive, which indicates that when the relationship between O-D station's land-use types is strong, the destination stations will be more likely to be chosen. The scale and collinear variable's parameters are positive, revealing that when the origin and destination stations' travel scale is more extensive, or the O-D station stands on the same line, the probability of the destination station being chosen is greater.

For the scenic variable in the one-way ticket model, its parameter is positive. It is also in line with the characteristics of passengers traveling on holidays because there are plenty of tourists who use the one-way tickets. In general, the estimated results are statistically significance and can explain the choice behavior mechanism on the New Year's Day to some degree. However, it is worth emphasizing that the parameters should be recalibrated so as to regain the travel behaviors when applying other different holidays.

4.2. Model Application and Comparison. To test the predictive effect of the proposed forecasting model, the calibrated results are used to predict the New Year's Day of Guangzhou Metro on January 1, 2017, where the data of the predicted year are used as the test-set and do not participate in the calibration. There are seventeen new stations and three new lines connected to the network. Meanwhile, the singly constrained gravity (SCG) model in the traditional statistical model, the support vector machine (SVM), the back propagation (BP) neural network, and radial basis function (RBF) neural network in machine learning model are selected for comparison under the same data source and conditions. And the traffic impedance function in the form of the exponential function is used in the gravity model, as shown in equation (13). The least-square method is used to transform it into a linear form for parameter estimation [35]:

$$f_{ij} = \exp\left(-\mu T_{ij} - \tau n_{ij}\right) T_{ij}^{-\gamma}, \tag{13}$$

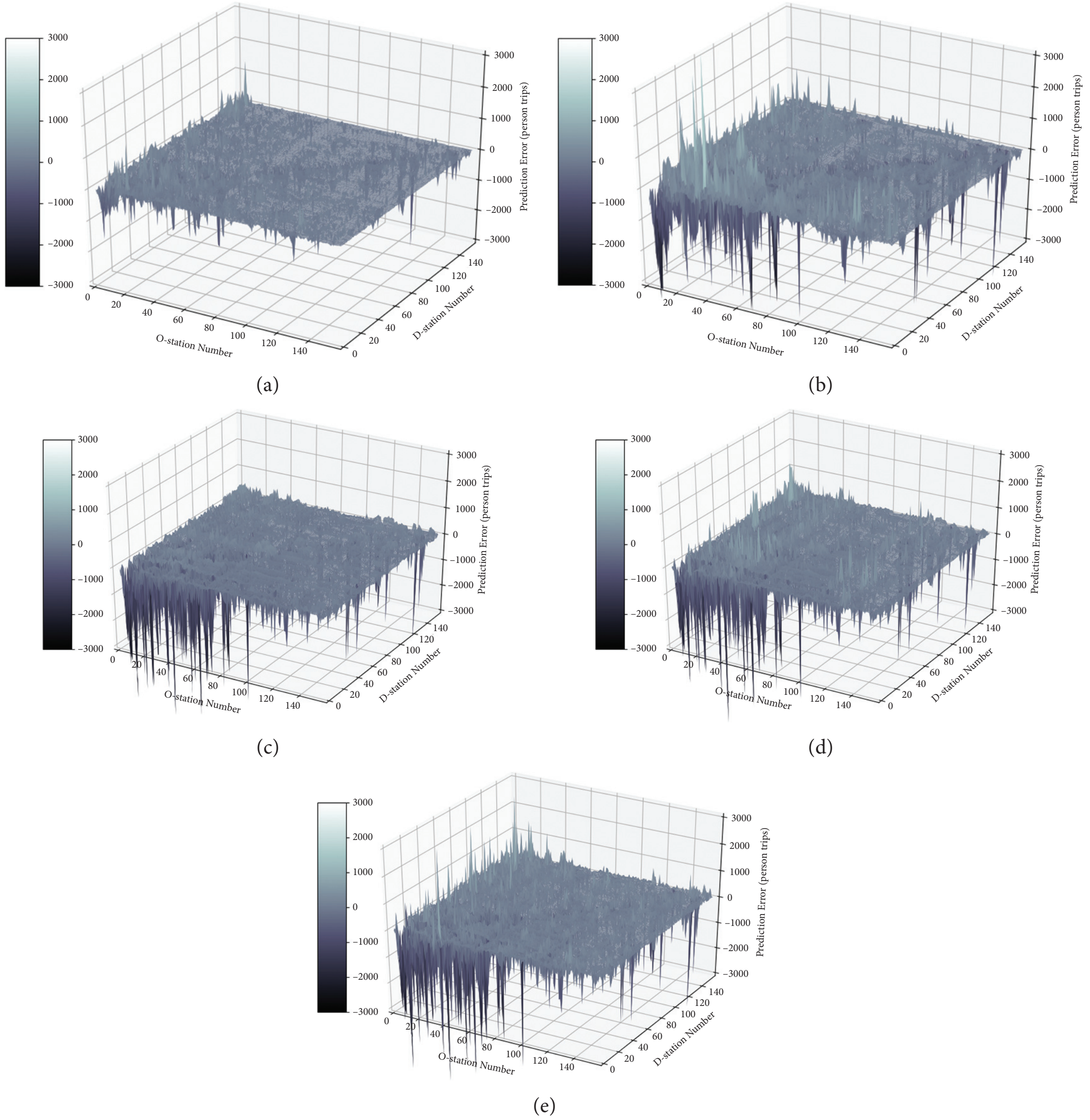
where T_{ij} and n_{ij} are travel time and transfer time from the origin station i to the destination station j, respectively; μ, τ, and γ are the coefficients to be determined.

As shown in Figure 4(a)–4(e), the predicted values are compared with the actual passenger travel data, and the prediction deviation graph is drawn. The error fluctuation of the singly constrained gravity model and the other three machine learning models is larger than the proposed forecasting model established in this paper. The mean absolute error of the whole network in the gravity model is 130.2 person trips, the SVM model is 140.9, the BP neural network is 139.1, and the RBF neural network is 157.3, while the proposed model is 54.6 that is far better.

Furthermore, the detailed prediction error statistics of the five models, in this case, are shown in Table 6. Compared with the other four models, the mean absolute error of the proposed model is reduced by 58.05%, 61.21%, 60.72%, and

Table 5: Model estimation results.

Characteristic variable		Case1: the New Year's Day				Case2: the National Day			
		One-way ticket		Public transport card		One-way ticket		Public transport card	
		β_i	t-value	β_i	t-value	β_i	t-value	β_i	t-value
Destination attraction (D_j)		0.301	37.989	0.084	26.818	0.290	38.791	0.211	28.646
Land matching degree (Z_{ij})		0.311	3.641	0.424	6.802	0.257	2.367	0.532	8.481
Collinear variable (S_{ij})		0.533	18.190	0.549	18.759	0.493	17.179	0.553	18.812
Scale variable (G_{ij})		0.638	29.106	0.533	23.835	0.629	29.519	0.517	23.696
Travel time (T_{ij}^{train})		−0.815	−18.286	−1.172	−25.162	−0.508	−11.884	−1.094	−23.556
Transfer time (N_{ij}^{trans})		−5.507	−18.324	−6.449	−20.833	−5.975	−20.316	−6.766	−21.519
Scenic variables (L_{ij})		0.084	3.971	—	—	0.092	3.350	—	—
Model summary	Observations	17954		18314		17534		17747	
	$L(\theta_{\max})$	−32640.37		−33441.36		−31654.83		−32615.35	
	$L(0)$	−41340.61		−42169.54		−40373.53		−40863.98	
	Adjusted ρ^2	0.210		0.207		0.216		0.202	



(a)

(b)

(c)

(d)

(e)

Figure 4: Prediction deviation of the proposed model and the comparison models on the New Year's Day. (a) The proposed model. (b) The gravity model. (c) The support vector machine model. (d) The back propagation neural network model. (e) The radial basis function neural network model.

Table 6: Statistics of model deviation in the New Year's Day.

Case1: the New Year's Day	SCG	SVM	BP	RBF	The proposed model
Maximum of absolute error (person trips)	5071	6099	6240	5932	2506
Mean absolute error (person trips)	130.22	140.86	139.08	157.27	54.63
Proportion of absolute errors over 50 trips (%)	50.13	56.12	57.57	66.3	26.90
Proportion of absolute errors over 100 trips (%)	31.31	31.83	35.71	39.29	14.78
Proportion of absolute errors over 200 trips (%)	16.27	15.17	14.74	17.15	5.90
Proportion of relative error over 20% (%)	84.60	85.11	84.00	85.95	62.74
Proportion of relative error over 50% (%)	66.78	64.34	64.08	68.37	33.17
Proportion of relative error over 100% (%)	40.45	37.55	45.70	46.07	16.39

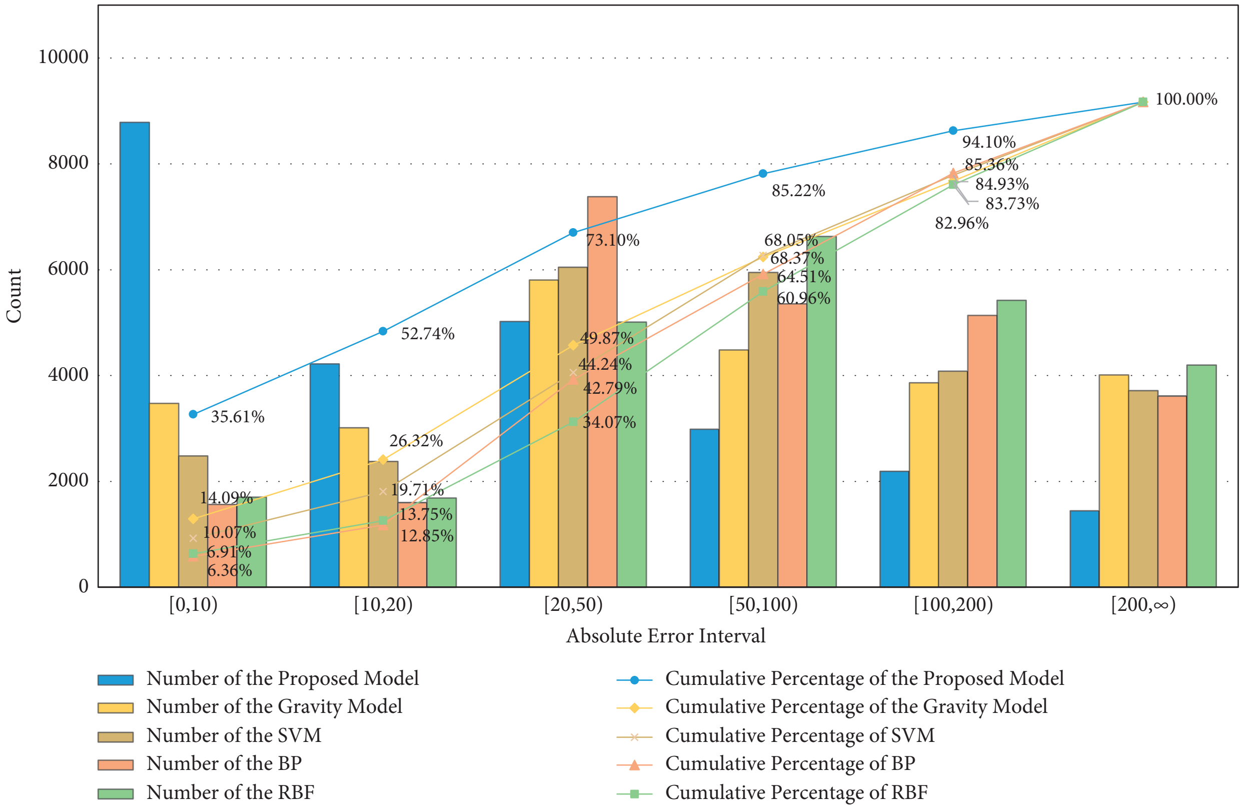


Figure 5: Comparison of absolute error and cumulative percentage.

65.26%, respectively. The proportion of absolute errors of the proposed model under 50 person trips reaches 73.1%, and the relative error less than 50% is 66.83%, where the errors are better than that of the other four models.

A detailed comparison of absolute error and its cumulative percentage can be seen in Figure 5. The statistics also show that the proposed model accuracy is better than the conventional gravity model, SVM, and the two neural network models as a whole. However, the proposed model has a slightly weak performance in terms of relative error, mainly because there are many O-D stations with small basic flow, leading to a large relative error. For example, the proportion of relative error more than 200% is 6.78%, where the average absolute error is 41.0 person trips. Moreover, the proportion of relative error more than 500% is 1.70%, where the average absolute error is 35.40 person trips, which is below the total average absolute error. Therefore, it does not mean that the poorer the relative error, the larger the absolute error, and the worse the prediction performance. The prediction effect of the proposed model can still be guaranteed.

Besides that, the error results of different categories between new lines and existing lines in the models are shown in Table 7. The proposed model's mean absolute error results are relatively low when predicting the new line, namely, only 23.14 and 23.26 person trips. In the prediction performance of the existing line to existing line, the error is relatively larger than that of others, mainly because of the large basic flow between existing stations.

In this case, the holiday of New Year's Day is chosen for analysis. However, other holidays might be a little longer in time, and passenger flow patterns and choice behavior would be different in some ways. The proposed destination choice model could be used to reflect the choice behavior

Table 7: Mean absolute error of prediction results.

Category	Mean absolute error (person trips)				
	SCG	SVM	BP	RBF	The proposed model
Existing to new lines	49.92	99.81	67.26	82.96	23.14
New line to existing line	48.07	76.62	119.80	105.64	23.26
New line to new line	90.03	141.72	96.56	131.55	41.30
Existing line to existing line	154.94	152.50	152.12	171.68	64.10
Whole network	**130.22**	**140.86**	**139.08**	**157.27**	**54.63**

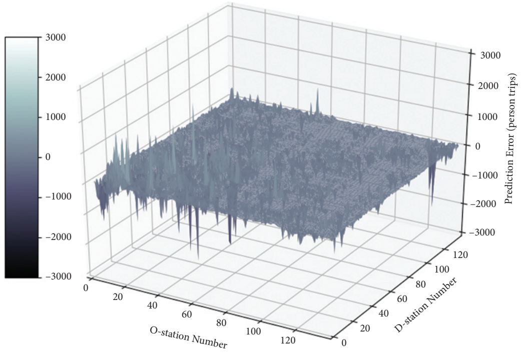


Figure 6: Prediction deviation of the proposed model on the National Day.

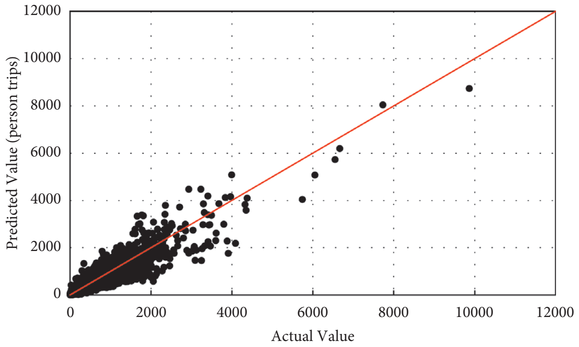


Figure 7: Scatter plot of predicted and actual value result presentation.

Table 8: Statistics of model deviation in the National Day.

Case2: the National Day	SCG	SVM	BP	RBF	The proposed model
Maximum of absolute error (person trips)	5681	8931	6730	4320	2467
Mean absolute error (person trips)	129.07	110.29	109.25	79.58	51.63
Proportion of absolute errors over 50 trips (%)	51.58	43.63	48.34	37.53	23.49
Proportion of absolute errors over 100 trips (%)	32.44	26.16	26.99	20.68	12.55
Proportion of absolute errors over 200 trips (%)	16.53	12.58	12.40	9.70	5.47
Proportion of relative error over 20% (%)	81.10	84.66	82.45	73.06	65.15
Proportion of relative error over 50% (%)	55.40	62.93	59.80	43.66	31.61
Proportion of relative error over 100% (%)	31.86	35.73	38.52	26.56	14.06

characteristics and passenger flow rules, so the methodology applies to all holidays. Considering the validation and portability of the proposed method, this study supplemented a case of the National Day (seven-day holiday) for a relatively comprehensive experimental design. One day of the National Day in 2014 was randomly selected for model estimation (that is, October 2, 2014), and the proposed method was used to predict the passenger flow distribution on the same day of next year.

There is one small difference in the calibrated parameters as shown in Table 5 above, which reflects the slight distinction of the travel characteristics in different holidays. However, all absolute t-values are still greater than 1.96, and the adjusted ρ^2 is over 0.2, which indicates that the model is still applicable and reliable. The prediction deviation graph is also drawn to show the overall error, as shown in Figure 6. For the convenience of reading, the scatter plot is given as shown in Figure 7, which is consistent with the meaning expressed in Figure 6. The graph plots values for the modeled prediction along the Y-axis and the corresponding actual count along the X-axis. If all of the predictions match the actual value exactly, the points on the graph would match up with the red line (45 degree line) drawn in the graph. The prediction results of the proposed model are mostly close to the red line, illustrating that the prediction performs well. The National Day's comparison models' error statistics are shown in Table 8. In summary, the prediction effect and accuracy are still ideal than the other models. The validation and applicability in other holiday scenarios can still be guaranteed. And it can be more effectively applied to practical engineering.

5. Conclusions

This paper utilizes AFC data to propose a forecasting model for passenger flow distribution for urban rail transit, which is suitable for network structure and the unique characteristics of holidays. The weighted exogenous sampling maximum likelihood (WESML) estimation method is used to calibrate the parameters. The aggregate data extracted from AFC are transformed into the disaggregate form, which realizes the valid calibration of the parameters. It reduces the difficulty of data acquisition and enhances the applicability of the model, meanwhile ensuring acceptable accuracy.

In the proposed model, the destination choice model defines destination attraction, land matching degree, and others as explanatory variables. This is the main advantage of the model's interpretability and predictive power. The model presents reasonable performance because t-values are all greater than 1.96, and the moderately adjusted ρ^2 is over 0.2. Moreover, the calibration results show that both travel and transfer time have significant negative effects on passengers' destination choice, while other variables such as destination attraction and land matching degree have a positive influence. The results also show that the public transportation card passengers care more about both travel and transfer time when other variables remain unchanged. The dummy variables used to describe the attractiveness and accessibility of the destination also have reasonable interpretability and significance. The proposed model is applied to predict two cases of Guangzhou Metro on New Year's Day and National Day. Compared with the gravity model, SVM, BP, and RBF models, the proposed model's error is greatly reduced, which proves the validation and applicability of the prediction model in different holiday scenarios with network changes.

As more cities rely on metro systems, accurately forecasted holiday passenger flow distribution could provide important primary data for the metro operation management department to develop a useful organization scheme before the holiday period, which is conducive to easing congestion and improving holiday emergency response capabilities.

Since it is difficult to obtain real land-use data around stations, this paper clusters the stations with similar passenger flow characteristics and defines new variables describing the land-use connection into the model. Nonetheless, the impact of significant land-use changes on passenger flow is hard to capture accurately. Furthermore, the dynamic characteristics of traffic flow distribution could be an extending study, which has not yet been considered in this paper. In future research, more land-use attributes and dynamic traffic distribution could be taken into account to develop the distribution forecasting model.

Data Availability

The data used to support the findings of this study were supplied by Guangzhou Metro under license and so cannot be made freely available. Access to these data should be considered by the corresponding author upon request, with permission of Guangzhou Metro.

Conflicts of Interest

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

Acknowledgments

This research was supported by the Fundamental Research Funds for the Central Universities (no. 2020YJS080) and the National Natural Science Foundation of China (no. 71931003).

References

[1] A. Salzberg, S. Mehndiratta, and Z. Liu, "Urban rail development in China," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2275, no. 1, pp. 49–57, 2012.

[2] C. Cai, E. Yao, S. Liu, Y. Zhang, and J. Liu, "Holiday destination choice behavior analysis based on AFC data of urban rail transit," *Discrete Dynamics in Nature and Society*, vol. 2015, Article ID 136010, 2015.

[3] S. Liu and E. Yao, "Holiday passenger flow forecasting based on the modified least-square support vector machine for the metro system," *Journal of Transportation Engineering, Part A: Systems*, vol. 143, no. 2, Article ID 04016005, 2016.

[4] L. He, Q. Liang, and S. Fang, "Challenges and innovative solutions in urban rail transit network operations and management: China's Guangzhou metro experience," *Urban Rail Transit*, vol. 2, no. 1, pp. 33–45, 2016.

[5] T. Grosche, F. Rothlauf, and A. Heinzl, "Gravity models for airline passenger volume estimation," *Journal of Air Transport Management*, vol. 13, no. 4, pp. 175–183, 2007.

[6] Y. Wang, L. Yang, Y. Geng, and M. Zheng, "OD matrix estimation for urban expressway," *Journal of Transportation Systems Engineering and Information Technology*, vol. 10, no. 2, pp. 83–87, 2010.

[7] M. Ren, Y. Lin, M. Jin, Z. Duan, Y. Gong, and Y. Liu, "Examining the effect of land-use function complementarity on intra-urban spatial interactions using metro smart card records," *Transportation*, vol. 47, no. 4, pp. 1607–1629, 2019.

[8] T. Tsekeris and A. Stathopoulos, "Gravity models for dynamic transport planning: development and implementation in urban networks," *Journal of Transport Geography*, vol. 14, no. 2, pp. 152–160, 2006.

[9] P. J. Jin, M. Cebelak, F. Yang, J. Zhang, C. M. Walton, and B. Ran, "Location-based social networking data," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2430, no. 1, pp. 72–82, 2014.

[10] A. Tsirimpa, A. Polydoropoulou, and C. Antoniou, "Development of a mixed multi-nomial logit model to capture the impact of information systems on travelers' switching behavior," *Journal of Intelligent Transportation Systems*, vol. 11, no. 2, pp. 79–89, 2007.

[11] Y. Yang, E. Yao, Z. Yang, and R. Zhang, "Modeling the charging and route choice behavior of BEV drivers," *Transportation Research Part C: Emerging Technologies*, vol. 65, pp. 190–204, 2016.

[12] D. Q. Nguyen-Phuoc, G. Currie, C. De Gruyter, and W. Young, "Transit user reactions to major service withdrawal - a behavioural study," *Transport Policy*, vol. 64, pp. 29–37, 2018.

[13] A. Faghih-Imani and N. Eluru, "Analysing bicycle-sharing system user destination choice preferences: chicago's Divvy system," *Journal of Transport Geography*, vol. 44, pp. 53–64, 2015.

[14] J. Kelly, A. Patrick, D. Christopher, and J. Robert, "Development of destination choice models for pedestrian travel," *Transportation Research Part A: Policy and Practice*, vol. 94, pp. 255–265, 2016.

[15] M. Mehadil Orvin and M. Rahman Fatmi, "Modeling destination choice behavior of the dockless bike sharing service users," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2674, no. 11, pp. 875–887, 2020.

[16] H. J. P. Timmermans, "A stated choice model of sequential mode and destination choice behaviour for shopping trips," *Environment and Planning A: Economy and Space*, vol. 28, no. 1, pp. 173–184, 1996.

[17] G. Jovicic and C. O. Hansen, "A passenger travel demand model for Copenhagen," *Transportation Research Part A: Policy and Practice*, vol. 37, no. 4, pp. 333–349, 2003.

[18] S. Ashiabor, H. Baik, and A. Trani, "Logit models for forecasting nationwide intercity travel demand in the United States," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2007, no. 1, pp. 1–12, 2007.

[19] A. Nuzzolo and A. Comi, "Urban freight demand forecasting: a mixed quantity/delivery/vehicle-based model," *Transportation Research Part E: Logistics and Transportation Review*, vol. 65, pp. 84–98, 2014.

[20] T. Li, "A demand estimator based on a nested logit model," *Transportation Science*, vol. 51, no. 3, pp. 918–930, 2017.

[21] P. Ye and D. Wen, "A study of destination selection model based on link flows," *IEEE Transactions on Intelligent Transportation Systems*, vol. 14, no. 1, pp. 428–437, 2013.

[22] M.-H. Wang, S. D. Schrock, N. V. Broek, and T. Mulinazzi, "Estimating dynamic origin-destination data and travel demand using cell phone network data," *International Journal of Intelligent Transportation Systems Research*, vol. 11, no. 2, pp. 76–86, 2013.

[23] Y. Wang, H. Yin, H. Chen, T. Wo, J. Xu, and K. Zheng, "Origin-destination matrix prediction via graph convolution: a new perspective of passenger demand modeling," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1227–1235, Anchorage, AK, USA, July 2019.

[24] K. E. Train, *Discrete Choice Methods with Simulation*, Cambridge University Press, Cambridge, UK, 2009.

[25] T. J. Fratar, "Vehicular trip distributions by successive approximations," *Traffic Quarterly*, vol. 8, no. 1, pp. 53–65, 1954.

[26] P. K. Sarkar, V. Maitri, and G. J. Joshi, *Transportation Planning: Principles, Practices and Policies*, PHI Learning Pvt. Ltd, New Delhi, India, 2017.

[27] A. K. Jain, "Data clustering: 50 years beyond *K*-means," *Pattern Recognition Letters*, vol. 31, no. 8, pp. 651–666, 2010.

[28] M.-K. Kim, S.-P. Kim, J. Heo, and H.-G. Sohn, "Ridership patterns at subway stations of Seoul capital area and characteristics of station influence area," *KSCE Journal of Civil Engineering*, vol. 21, no. 3, pp. 964–975, 2017.

[29] X. Zhao, Y. Wu, G. Ren, K. Ji, and W. Qian, "Clustering analysis of ridership patterns at subway stations: a case in Nanjing, China," *Journal of Urban Planning and Development*, vol. 145, no. 2, Article ID 04019005, 2019.

[30] E. Yao and T. Morikawa, "A study of on integrated intercity travel demand model," *Transportation Research Part A: Policy and Practice*, vol. 39, no. 4, pp. 367–381, 2005.

[31] C. F. Manski and S. R. Lerman, "The estimation of choice probabilities from choice based samples," *Econometrica*, vol. 45, no. 8, pp. 1977–1988, 1977.

[32] S. R. Cosslett, "Efficient estimation of discrete-choice models," *Structural Analysis of Discrete Data with Econometric Applications*, vol. 3, pp. 51–111, 1981.

[33] M. E. Ben-Akiva and S. R. Lerman, *Discrete Choice Analysis: Theory and Application to Travel Demand*, MIT Press, Cambridge, MA, USA, 1985.

[34] D. McFadden, *Quantitative Methods for Analyzing Travel Behaviour of Individuals: Some Recent Developments*, Cowles Foundation for Research in Economics, New Haven, CT, USA, 1977.

[35] M. Nakanishi and L. G. Cooper, "Parameter estimation for a multiplicative competitive interaction model-least squares approach," *Journal of Marketing Research*, vol. 11, no. 3, pp. 303–311, 1974.

Hindawi
Journal of Advanced Transportation
Volume 2021, Article ID 7763126, 15 pages
https://doi.org/10.1155/2021/7763126

Research Article

Prediction of Train Arrival Delay Using Hybrid ELM-PSO Approach

Xu Bao,[1] **Yanqiu Li**,[2] **Jianmin Li**,[2] **Rui Shi**,[2] **and Xin Ding**[2]

[1]*College of Traffic Engineering, Huaiyin Institute of Technology, Jiangsu 223001, China*
[2]*School of Traffic and Transportation, Beijing Jiaotong University, Beijing 100044, China*

Correspondence should be addressed to Xu Bao; baoxu@hyit.edu.cn

Received 10 May 2021; Accepted 4 June 2021; Published 14 June 2021

Academic Editor: Linchao Li

In this study, a hybrid method combining extreme learning machine (ELM) and particle swarm optimization (PSO) is proposed to forecast train arrival delays that can be used for later delay management and timetable optimization. First, nine characteristics (e.g., buffer time, the train number, and station code) associated with train arrival delays are chosen and analyzed using extra trees classifier. Next, an ELM with one hidden layer is developed to predict train arrival delays by considering these characteristics mentioned before as input features. Furthermore, the PSO algorithm is chosen to optimize the hyperparameter of the ELM compared to Bayesian optimization and genetic algorithm solving the arduousness problem of manual regulating. Finally, a case is studied to confirm the advantage of the proposed model. Contrasted to four baseline models (k-nearest neighbor, categorical boosting, Lasso, and gradient boosting decision tree) across different metrics, the proposed model is demonstrated to be proficient and achieve the highest prediction accuracy. In addition, through a detailed analysis of the prediction error, it is found that our model possesses good robustness and correctness.

1. Introduction

With the rapid development of society and the continuous improvement of people's quality of life, people have put forward higher requirements for the reliability and punctuality of high-speed railway transportation [1]. However, the train will inevitably be disturbed by a large number of random factors in the process of running, which will lead to the train delay. For one thing, train delay will change the structure of train diagram, increase the cost of railway operation and the difficulty of reasonable utilization of transportation resources, and have a great negative impact on the reliability and punctuality of high-speed railway operation. For another, it will increase the travel time of passengers, affect their travel plans, and bring serious inconvenience to passengers [2]. Therefore, accurate forecast of train delay is of great significance for high-speed train operation organization, transportation service quality improvement, and operation safety [3].

The traditional models are a classical approach for train delay prediction, such as probability distribution models [4, 5], regression models, event-driven methods, and graph theory-based approaches. For the probability distribution model, Higgins and Kozan proposed an exponential distribution model, which applied a three-way, two-block station train delay propagation signal system, to estimate delays of trains caused by train operational accidents [4]. Through the assessment of the linear relationship between several independent features and dependent features [2], regression models were widely employed to predict train delays, dwell times, and running times [6, 7]. However, the main drawback of regression models is that the ability of linear analytic model relies much on internal and mathematical assumptions. They are good at capturing the linear relationship between features and dealing with low-dimensional data, not simple linear data, such as train operation data [8]. For event-driven and graph theory-based methods, Kecman and Goverde [9] used timed event graph with dynamic weight to predict train running time. Milinković et al. [10] used fuzzy Petri net to predict the train delays; the model considers the characteristics of the hierarchical structure and fuzzy reasoning to simulate the train operation and predicts the train delays in different delay scenarios. Huang et al. [8] used graph theory to calculate the degree and propagation range of train delay under specific condition. Although the traditional model started early in the study of delay prediction, it generally has the limitation of poor generalization performance and is only suitable for specific scenarios.

Recently, the application of machine learning methods to predict train delays has been widely concerned by researchers, which makes up for the shortcomings of traditional methods [2]. The purpose of Peters et al. was to utilize the historical travel time between stations to predict the train arrival time more precisely [11]. The moving average algorithm of historical travel time and KNNs of last arrival time algorithm were employed and estimated. Some researchers are devoted to ANNs to predict train delays [12–14]. The aim of [14] was to propose preeminent ANNs to predict the train delay of Iran railway with three different models, including standard real number, binary coding, and binary set encoding inputs. Nevertheless, the prediction accuracy of ANNs cannot meet the needs of actual delay management, and the parameter adjustment is complex. Marković et al. [2] proposed a support vector regression model in train delay problem of passenger train, which captured the relationship between the arrival delay and a variety of changing external factors, and compared it with the artificial neural networks. The results indicated that the support vector regression method outperformed the ANNs. Another neural network has been proposed in recent years. A Bayesian network model for predicting the propagation of train delays was presented by [15]. In view of the complexity and dependence, three different BN schemes for train delay prediction were proposed, namely, heuristic hill-climbing, primitive linear, and hybrid structures [16]. The results turned out to be quite satisfying. Recently, it has become popular to combine several models to capture various characteristics of train operation data to predict train delay. A study developed a train delay prediction model, which combines convolutional neural networks, long short-term memory network, and fully-connected neural network architectures to solve this issue [17].

To improve the backpropagation algorithm and simplify the setting of learning parameters of general machine learning models, the ELM algorithm was proposed by Bin Huang et al. [18]. ELM has the advantages of small computation, good generalization, and fast convergence. On account of these advantages, ELM has been frequently applied to regression problems in the real world [19–22]. Therefore, a new study that combined a shallow ELM and a deep ELM tuned via the threshold out technique was employed to predict train delays, taking the weather data into account [23].

Parameter adjustment is another critical factor to guarantee the good performance of machine learning models [24, 25]. Although the well-known random search algorithm can achieve the purpose of optimization, it generates all the solutions randomly without considering the previous solutions. An adjusting parameter model, called PSO, has become one of the widely used parameter adjustment methods because of its ability to address intractable matters in the real world. Only the optimal particle of PSO transmits the information to the next particle in the iterative evolution process. As a consequence, the searching speed of PSO is faster than random search and grid search [26]. The experiment [27] did just prove the advantage of PSO. By comparing the performance of PSO with random search algorithm for the optimal control problem, [27] found out that PSO was capable of locating better solution with the same number of fitness function calculations than random search algorithm.

Therefore, according to what the author has learnt, we propose PSO to optimize the hyperparameter of ELM to forecast train arrival delays.

The contributions this paper makes are as follows:

(1) The main features affecting the train delay prediction are evaluated by the extra trees classifier. Then, the proposed model is constructed based on these features which possess spatiotemporal characteristics (train delays at each station). In this way, the interpretability of the proposed model is improved.

(2) The proposed model is applied to the arrival delay prediction of trains on HSR line, which suggests a brand-new perspective for the train delay prediction problem. In addition to solving the drawbacks of backpropagation algorithm, the advantage of ELM-PSO is also to solve the arduous problem of manual regulating the hidden neurons of ELM better than random search and Bayesian optimization at accuracy and efficiency.

(3) We perform experiments on a section of the Wuhan-Guangzhou (W-G) HSR line. The proposed model not only is compared to other two adjusting parameter models, but also is contrasted with four prediction models from different perspectives. Our model turns out to have an extraordinary ability in managing large-scale data in accuracy.

The remainder of this paper is distributed as follows: in Section 2, the train delay problem and selection of characteristic features are described. The hybrid ELM-PSO approach is introduced in detail in Section 3. The data description and experimental settings are presented in Section 4. The performance analysis is discussed in Section 5. Finally, conclusions are presented in Section 6.

2. Description of the Train Delay Problem

Train delay problem is visualized in Figure 1 to assist in comprehending this abstract problem. The train delay contains two contents, train arrival delay and train departure delay. For a station s_n, t_{Asn} represents the time that one train is scheduled to arrive at station s_n and the same goes for t_{Dsn}, which implies the time that one train is scheduled to depart at station s_n. Certainly, the train will have its own actual timetable due to changing external factors, which are

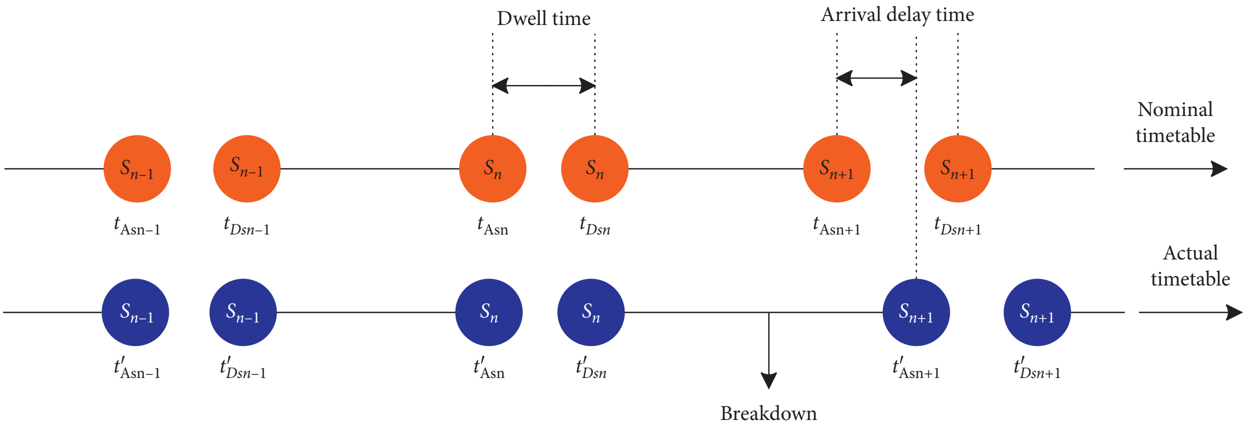


Figure 1: Conversion from the train itinerary to mathematical notation.

expressed as t'_{Asn} and t'_{Dsn}, respectively. The difference between the actual and scheduled arrival time at station s_n, $t'_{\mathrm{Asn}} - t_{\mathrm{Asn}}$, is referred to as the train arrival delay. The same goes for the train departure delay $t'_{\mathrm{Dsn}} - t_{\mathrm{Dsn}}$. This is the primitive description of the train delay problem.

This paper only focuses on the train arrival delay prediction. We suppose that there is an aimed train T_k, which is at present station s_n at time t_{Asn}. Our purpose is to predict the arrival delay ($t'_{\mathrm{Asn}+1} - t_{\mathrm{Asn}+1}$) of the targeted train $\mathrm{T}_k$at its following station s_{n+1} for all conditions according to the information of train k at stations s_n, s_{n+1}, and s_{n-1}, which is made up of the following nine features:

(1) The station code (X_1)

(2) The train number (X_2), which indicates the number of the trains

(3) The length between the present station and the next station ($\mathrm{D}_{\mathrm{sn}+1} - D_{\mathrm{sn}}$) ($X_3$)

(4) The scheduled running times between the present station and the previous station ($t_{\mathrm{Asn}} - t_{\mathrm{Asn}-1}$) ($X_4$)

(5) The actual running times between the present station and the previous station ($t'_{\mathrm{Asn}} - t'_{\mathrm{Asn}-1}$) ($X_5$)

(6) The scheduled running times between the present station and the next station ($t_{\mathrm{Asn}+1} - t_{\mathrm{Asn}}$) ($X_6$)

(7) The actual running times between the present station and the next station ($t'_{\mathrm{Asn}+1} - t'_{\mathrm{Asn}}$) ($X_7$)

(8) Buffer time, which indicates the difference between X_6 and actual minimum running time of all trains between the present station and the next station ($X_6 - \min\{T_1(t'_{\mathrm{Asn}+1} - t'_{\mathrm{Asn}}) \cdots T_k(t'_{\mathrm{Asn}+1} - t'_{\mathrm{Asn}})\}$) ($X_8$)

(9) The arrival delay time at the present station ($t'_{\mathrm{Asn}} - t_{\mathrm{Asn}}$) ($X_9$)

Y represents the arrival delay time at the next station s_{n+1} of train T_k ($t'_{\mathrm{Asn}+1} - t_{\mathrm{Asn}+1}$).

There are multiple potential interdependent features (e.g., the train number, the length between two adjacent stations) that are intently related to train delay prediction. Based on the collected data and the experience of dispatchers, we ultimately select nine features that are possible to influence train delays.

We apply extra trees classifier to analyze the correlation between all features and train delays. The results are exhibited in Figure 2. As shown in the figure, the deeper the red, the higher the importance. There is no doubt that X_9 has the highest importance with Y. The actual and scheduled running times between the present station and the next station also contribute largely to the accurate prediction of Y. Moreover, the buffer time, which is an important factor affecting the length of the train recovery time, is also comparatively prominent in delay prediction process. Taking the buffer time into account allows us to obtain more realistic prediction results.

The train arrival delay prediction problem in this paper is transformed into the following expression:

$$Y = f(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9), \qquad (1)$$

where X_i is the information of train T_k running through stations s_n, s_{n+1}, and s_{n-1}, Y is the arrival delay time at the following station s_{n+1} of train T_k, and $f(x)$ is the prediction process.

3. Methodology

This paper proposes a hybrid model of ELM and PSO for train delay prediction. ELM is widely used in regression problems because of its advantages of small computation, good generalization performance, and fast convergence speed [19–21]. PSO algorithm is a random and parallel optimization algorithm, which has the advantages of fast convergence speed and simple algorithm [25, 28]. Therefore, we aim to combine the advantages of ELM and PSO algorithm to improve the behavioral knowledge in the delay prediction domain. For the principle of ELM and PSO, one can refer to Li et al. [29], Perceptron et al. [30], and Zhang et al. [31]. The running process of the proposed hybrid method is as follows:

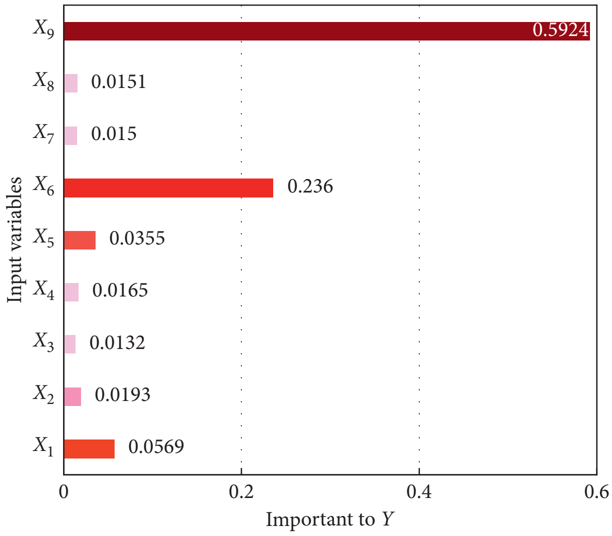


Figure 2: Bar chart of the correlations between the nine input features and output Y.

Step 1: data preprocessing. First, 9 features mentioned in Section 2 are generated a $N \times 9$ matrix, where N represents the total number of events according to the train operation records. Second, remove abnormal delay (trains may be canceled due to some emergencies) to reduce its interference with predictions. Third, fill in the missing data according to the adjacent data around the missing ones.

Step 2: initializing the parameters and population. Parameters such as maximal iteration number, population size, and speed and position of the first particle are initialized. Each particle l has its own position $A_l(ite)$ and speed $V_l(ite)$. The position of each particle in the population is equivalent to the number of neurons in the hidden layer of ELM. Therefore, there is merely one dimension of each particle:

$$A_l(\text{ite}) = h_l(\text{ite}), \quad l = 1, \ldots, k, \tag{2}$$

where $h_l(\text{ite})$ represents the number of hidden layer neurons of ELM in the ite_{th} iteration.

Step 3: ELM (hidden layer activation function: sigmoid function) is used. The processed feature set X and the position of particles (the number of hidden layer neurons) generated by PSO are input into ELM. Consequently, ELM can output the weight matrix under the current number of hidden layer neurons. The function of calculating the fitness of particles is as follows:

$$p_l(\text{ite}) = \text{fitness}(A_l(\text{ite}), X) = \sqrt{\frac{\sum_{n=1}^{N}(y_n - f_n(x))^2}{N}},$$
$$A_l(\text{ite}) = h_l(\text{ite}), \quad l = 1, \ldots, k, \tag{3}$$

where N is the number of samples, y_n is the actual output value on test set, and $f_n(x)$ is the predicted output value on test set.

Step 4: calculate the fitness of each particle, and compare to update the current best fitness and its particle location.

Step 5: start the iteration. PSO will update the positions and velocities of all particles, and then repeat step 4. If the maximum number of iterations is exceeded, it will end the process.

Step 6: output the results. We can obtain the output value on test set as well as the optimal number of hidden layer neurons.

The specific flowchart is shown in Figure 3.

4. Application to a Case Study

4.1. Dataset Description. The data employed to verify the ELM-PSO are obtained from the dispatching office of a railway bureau. The 15 stations applied in the study include a section, the length of which is 1096 km from CBN to GZS on the double-track W-G HSR line. There are more than 400,000 data points used in this study, with a time span from October 2018 to April 2019. The train original operation data and route map of the targeted 15 stations on the W-G HSR line are shown in Table 1 and Figure 4.

Analysis of the delay ratio of each station reveals not only the condition of each station but also an increasing emphasis on the indispensability of train arrival delay prediction, which contributes to improving the ability of each station to cope with and even inhibit the increase in train arrival delays. Trains with arrival delay greater than 4 minutes are considered as delayed trains. What is intuitively presented in Figure 5 is that the delay ratios of all the stations are basically

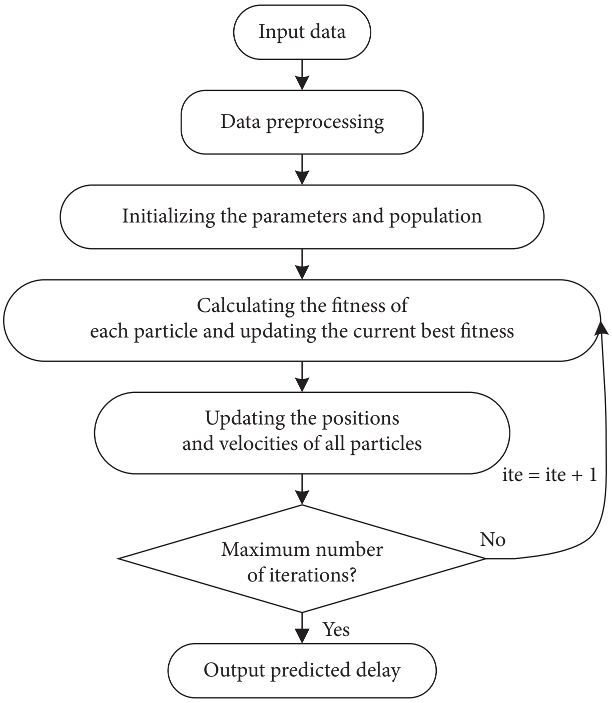

Input data
Data preprocessing
Initializing the parameters and population
Calculating the fitness of each particle and updating the current best fitness
Updating the positions and velocities of all particles
ite = ite + 1
No
Maximum number of iterations?
Yes
Output predicted delay

Figure 3: The flowchart of ELM-PSO method.

Table 1: Train operation data format in the database.

Station	Station code	Date	Actual arrival	Actual departure	Train	Scheduled arrival	Scheduled departure
GZS	369	2018/7/27	12:04	12:04	G100	12:05	12:05
GZN	368	2018/7/27	12:19	12:19	G100	12:19	12:19
QY	367	2018/7/27	12:26	12:26	G100	12:26	12:26
YDW	366	2018/7/27	12:37	12:37	G100	12:38	12:38

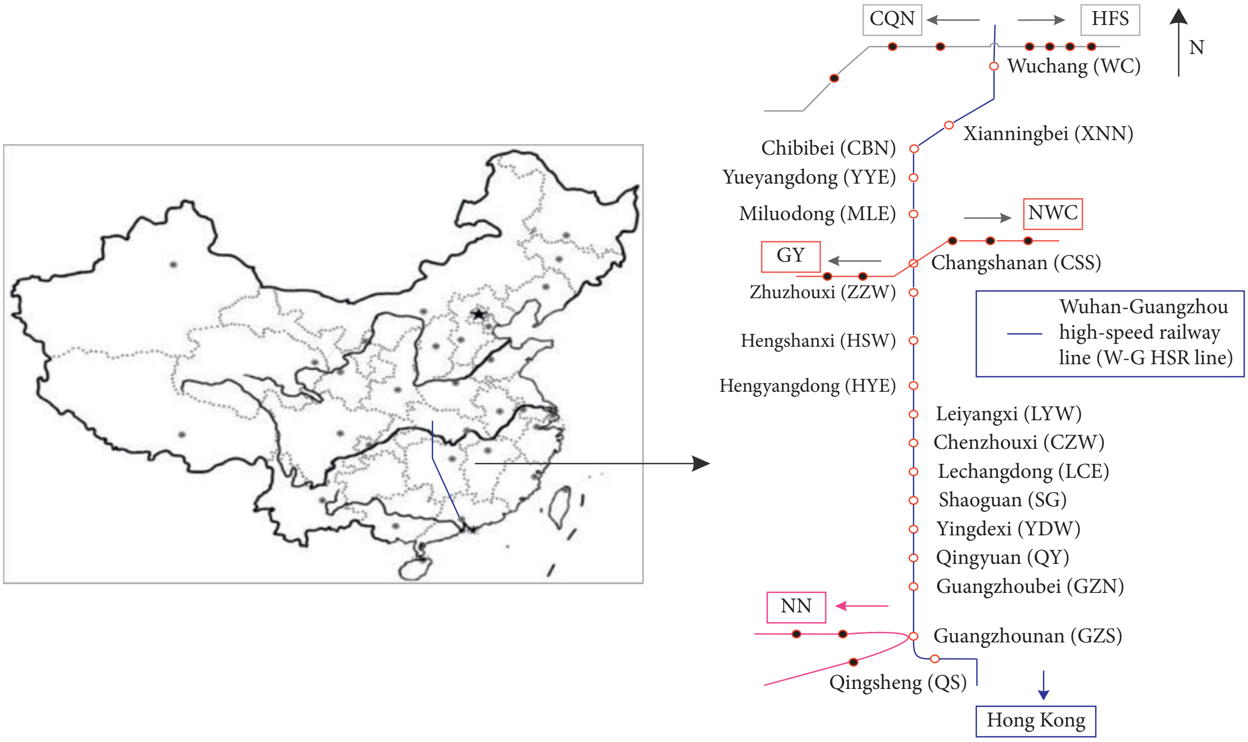


Figure 4: Map of the W-G HSR line.

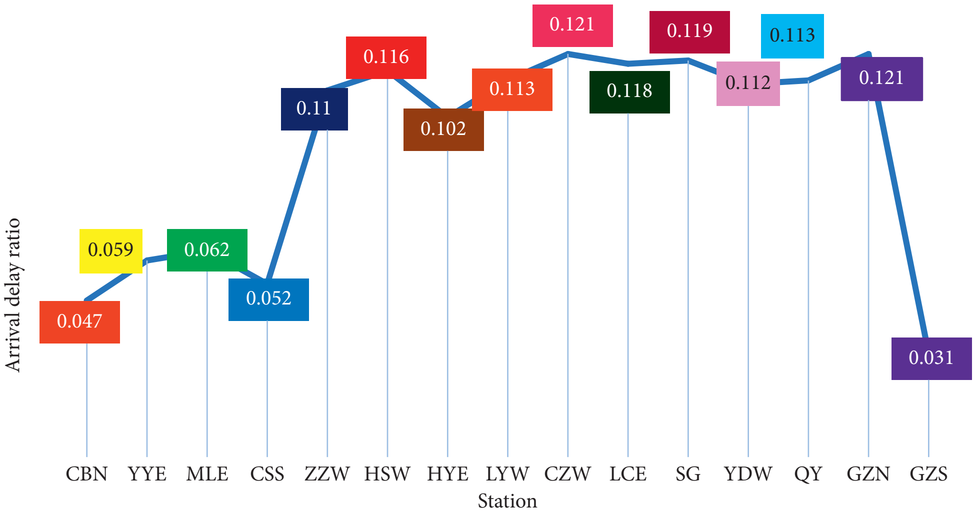


Figure 5: Arrival delay ratio for each station.

not optimistic. At the same time, the delay ratios of the two targeted stations, CZW and GZN, are particularly dreadful, with arrival delay ratios of 0.12. Our goal is to minimize the arrival delay ratio by predicting the arrival delay at each station.

4.2. Experimental Settings

4.2.1. Baseline Models. In order to compare the performance of our proposed method, the k-nearest neighbor (KNN), categorical boosting (CB), gradient boosting decision tree (GBDT), and Lasso are used as baseline models. We take 20% of the dataset as the test set and the rest as the training set. The experiment runs in Python in an environment with an Intel® Core i5-6200U processor 2.13 GHz and 8 GB RAM. Briefly, an overview description and hyperparameter settings of each model are as follows:

(1) KNN: KNN algorithm is extensively applied in differing applications massively, owing to its simplicity, comprehensibility, and relatively promising manifestation [32].

① N_neighbors = 15
② Weights = uniform
③ Leaf_size = 30
④ $P = 2$

(2) CB: CB is a machine learning model based on gradient boosting decision tree (GBDT) [33, 34]. CB is an outstanding technology, especially for datasets with heterogeneous features, noisy data, and complex dependencies.

① Depth = 3
② Learning_rate = 0.1
③ Loss_function = RMSE

(3) GBDT: GBDT has been employed to numerous problems [35], which has many nonlinear transformations and strong expression ability and does not need to do complex feature engineering and feature transformation.

① N_estimators = 30
② Loss = ls
③ Learning_rate = 0.1

(4) Lasso: Lasso is a prevailing technique, capable of simultaneously performing regularization and feature filtering. Furthermore, data can be analyzed from multiple dimensions by Lasso [36].

① Alpha = 3.0.
② Max_iter = 1000.
③ Selection = cyclic.

4.2.2. Evaluation Metrics. Root mean squared error (RMSE), mean absolute error (MAE), and *R*-squared are selected to assess the models. The definitions of the error metrics are shown in equation (4), equation (5), and equation (6):

$$\mathrm{MAE} = \frac{\sum_{i=1}^{N} \left| y_i - \hat{y}_i \right|}{N}, \tag{4}$$

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{N} \left(y_i - \hat{y}_i \right)^2}{N}}, \tag{5}$$

$$R - \mathrm{squared} = 1 - \frac{\sum_{i=1}^{N} \left(y_i - \hat{y}_i \right)^2}{\sum_{i=1}^{N} \left(y_i - \overline{y}_i \right)^2}, \tag{6}$$

where y_i is an observed value, $\hat{y}_i$ is a predicted value, $\overline{y}_i$ is the average value of y_i, and N represents the sample size.

4.2.3. Hyperparameter Tuning Models. We compare PSO with the other two hyperparameter tuning models to ascertain the most satisfying one. The overview and hyperparameter settings of each model are as follows:

(1) PSO: to locate the optimal hyperparameter of the ELM, the parameter settings of the PSO algorithm are as follows. PSO has 20 particles at each iteration, and there are altogether 20 iterations, which is equivalent to 400 iterations of Bayesian optimization.

① Number of particles = 20
② Fitness function: RMSE on test set
③ Search dimension = 1
④ Particle search range = [1, 2000]
⑤ Maximum number of iterations = 20

(2) BO (Bayesian optimization): BO calculates the posterior probability distribution of the first n points through a substitution function and obtains the objective function of each hyperparameter at each value point.

① Objective function: RMSE on test set
② Substitution function: Gaussian process regression
③ Acquisition function = UCB (upper confidence bound)
④ Hyperparameter search range = [1, 2000]
⑤ Maximum number of iterations = 400

(3) GA (genetic algorithm): the traditional iterative model is easy to fall into the trap of local minima, which makes the iteration impossible to continue. GA overcomes the phenomenon of "dead loop" and is a global optimization algorithm [37].

① Objective function: RMSE on test set
② Hyperparameter search range = [1, 2000]
③ Generations = 20
④ Population size = 20
⑤ Maximum number of iterations = 400

5. Performance Analysis

5.1. PSO Optimization Result Comparison. The process of PSO tuning the hyperparameter is shown in Figure 6. The fitness value achieves minimum after five iterations. The best fitness value is 1.0387 on test set when there are 1462 neurons of the ELM. The structure of the network is optimal correspondingly.

The search range [1–2000] of hyperparameter is determined by manually trying several values in the range of [1–10000]. When the hyperparameter value is greater than 2000, the fitness tends to be stable. Also, the time consumption is multiplied acutely. Ultimately, we decide to limit the search range to [1–2000], weighing time consumption and precision.

The computational cost is shown in Table 2, and the results are the optimal results of each model running several times. We gain two observations from this table. First, the optimal particle number of ELM-PSO always focuses on 1462; the only difference is the number of iterations at best RMSE. Second, compared with ELM-BO and ELM-GA, ELM-PSO is the ideal model that takes the shortest time to locate the optimal fitness on the test set.

5.2. Model Accuracy Comparison. In this section, the performance comparison between ELM-PSO and baseline models is performed.

First, we compare the overall performance of the five models. The evaluation metrics are R-squared, the MAE, and the RMSE. The corresponding results on test set and training set are summarized in Tables 3 and 4, respectively. The ELM-PSO model performs optimally among the five models in not only the training set (R-squared = 0.9973; MAE = 0.3377; RMSE = 0.8247), but also the test set (R-squared = 0.9955; MAE = 0.3490; RMSE = 1.0387). Although the running time of our model has no obvious advantage compared with other models on test set, it is within the tolerable range. Also, we notice that there are models that perform well in the training set, but are not good in the test set, which reveals the paramountcy of enough generalization ability of models in the prediction problem.

Then, by separating the delay duration into three bins (i.e., [0–1200 s], >1200 s, and all delayed trains (trains with arrival delay greater than 240 seconds)), we attempt to measure the capability of the benchmark models and our model to seize the features of train delays to varying degrees on test set. As is distinctly shown in Table 5, the proposed model outperforms the other benchmark models in each time horizon and each evaluation metric, achieving, for example, an RMSE of 0.5201 in the first bin. This finding is taken as evidence that our model can constantly adjust itself to capture the characteristics of varying degrees of train delays to enhance the prediction accuracy. To further assess the performance of our model, the comprehensive analyses are discussed in a later section.

5.3. Further Analysis. On the basis of the previous section, we will evaluate the performance of the ELM-PSO model from other angles, including the prediction errors for each station precisely, the prediction correctness, and the robustness.

First and foremost, the errors of the ELM-PSO model for the predicted arrival delays are calculated at the station level on test set. Viewing the overall situation in Figure 7, we have noticed that the prediction errors are low. The MAE and R-squared both remain stable at each station. And the RMSEs for different stations are mostly less than 90s. However, great fluctuations occur at the YYE and QY stations. Reasons resulting in such phenomena are that the two stations are close to the transfer stations and the buffer times of YYE and QY are both small. The prediction accuracy at these stations tends to be slightly hindered by these factors.

In addition, to put forward more detailed and embedded results, we describe the correctness of the absolute residual between the predicted values and the actual values for each station from three intervals (i.e., <30 s, 30 s–60 s, and 60 s–90 s) (Figure 8). In <30 s interval, the correctness of

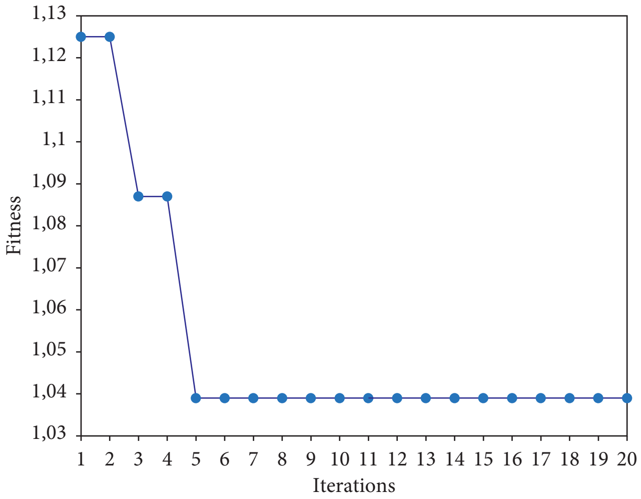


Figure 6: Convergence discriminant graph of PSO optimization.

Table 2: The performance of each hyperparameter setting model.

Model	Maximum number of iterations	The number of iterations at best RMSE	RMSE	Neurons	Time (s)
ELM-PSO	20	5	1.0387	1462	90000
ELM-BO	400	120	1.0708	1279	129600
ELM-GA	20	6	1.0668	1494	108000

Table 3: Prediction errors on each model's test set for the W-G HSR line.

Model	RMSE	MAE	*R*-squared	Time (s)
ELM-PSO	1.0387	0.3490	0.9955	856
CB	1.6808	1.0464	0.9883	9
GBDT	1.9976	1.2031	0.9835	18
Lasso	1.9852	0.9240	0.9847	9
KNN	1.6488	0.5025	0.9887	27

Table 4: Prediction errors on each model's training set for the W-G HSR line.

Model	RMSE	MAE	*R*-squared
ELM-PSO	0.8247	0.3377	0.9973
CB	1.6368	1.0407	0.9896
GBDT	1.9495	1.2107	0.9852
Lasso	1.6046	0.9199	0.9893
KNN	1.4681	0.4558	0.9916

each stations exceeds 75%. In brief, the overall results confirm the impressive prediction correctness of the proposed model.

At last, we investigate the robustness of our model to data size. In detail, we further train and test our model using 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, and 90%, respectively, of the total data as test set, and compare the results with the baseline models. The data sizes used in the experiments are shown in Figure 9. The performance of our model on both training and test set is more outstanding than others. As we can see, the RMSEs of our model stay pretty stable using data with different sizes, while the RMSEs of other models are higher and fluctuating. These figures show that the proposed model has the smallest predictive RMSE, MAE, and *R*-squared for all trains, which demonstrates the robustness of our model to different data sizes.

5.4. Statistical Tests. In this section, the Friedman test (FT) and Wilcoxon signed rank test (WSRT) are used to verify the advantages of our proposed method compared with other methods [35, 38]. The results FT and WSRT are shown in Table 6. FT algorithm is a nonparametric statistical tool, which determines the difference by ranking

Table 5: Model performance comparison on test set for the five models for different delay bins.

Delay bin (seconds)	Model	RMSE	MAE	*R*-squared
[0, 1200]	ELM-PSO	0.5201	0.3006	0.9655
	CB	1.2628	0.9223	0.7967
	GBDT	1.3968	0.9886	0.7513
	Lasso	1.0957	0.7903	0.8469
	KNN	1.0195	0.5278	0.8675
>1200	ELM-PSO	5.6009	2.8249	0.9924
	CB	6.2335	3.1986	0.9906
	GBDT	8.0733	4.8589	0.9843
	Lasso	6.5023	3.1169	0.9898
	KNN	9.0247	4.8611	0.9804
All delayed trains (>240)	ELM-PSO	3.3116	1.3457	0.9951
	CB	4.0469	2.2680	0.9927
	GBDT	5.1561	3.1498	0.9881
	Lasso	3.9251	1.7418	0.9931
	KNN	5.5878	2.8828	0.9860

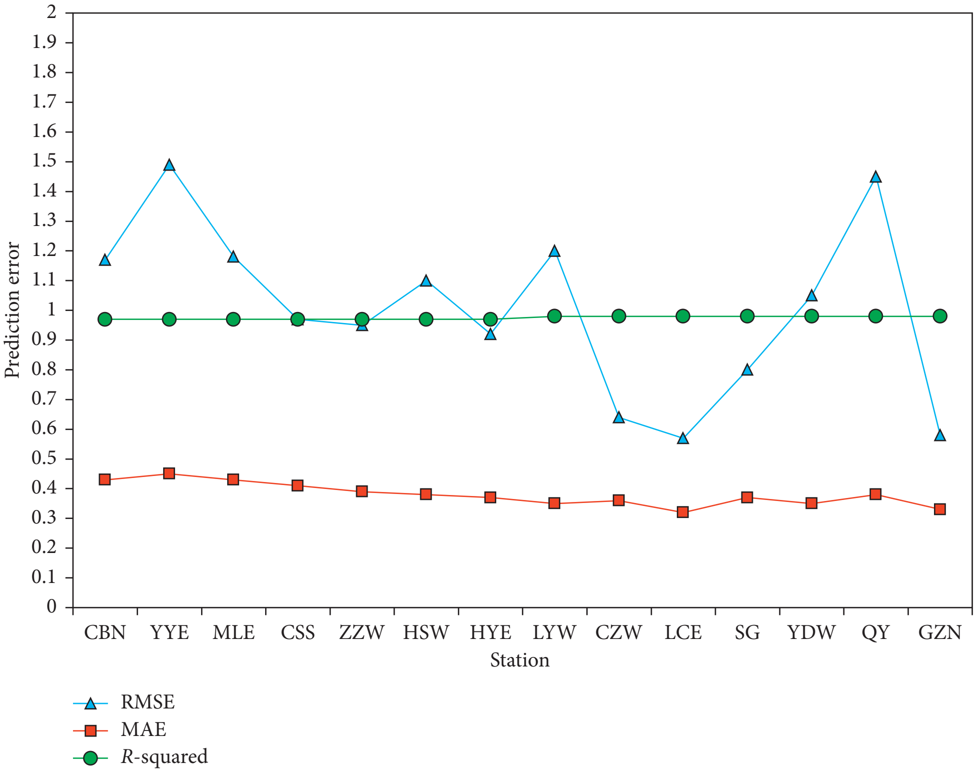


Figure 7: Prediction errors in terms of the RMSE, MAE, and *R*-squared for different stations.

the performance of each method. It can be seen from the table that the proposed method has better ranking than CB, GBDT, Lasso, and KNN at 5% significance level; that is, the efficiency is better. In addition, the results of WSRT showed that the *p*-value was less than 0.05 (5% significance level), which rejected the null hypothesis. It means that there is a statistical difference between the proposed method and other methods. That is, the performance of the proposed method is better than that of other methods.

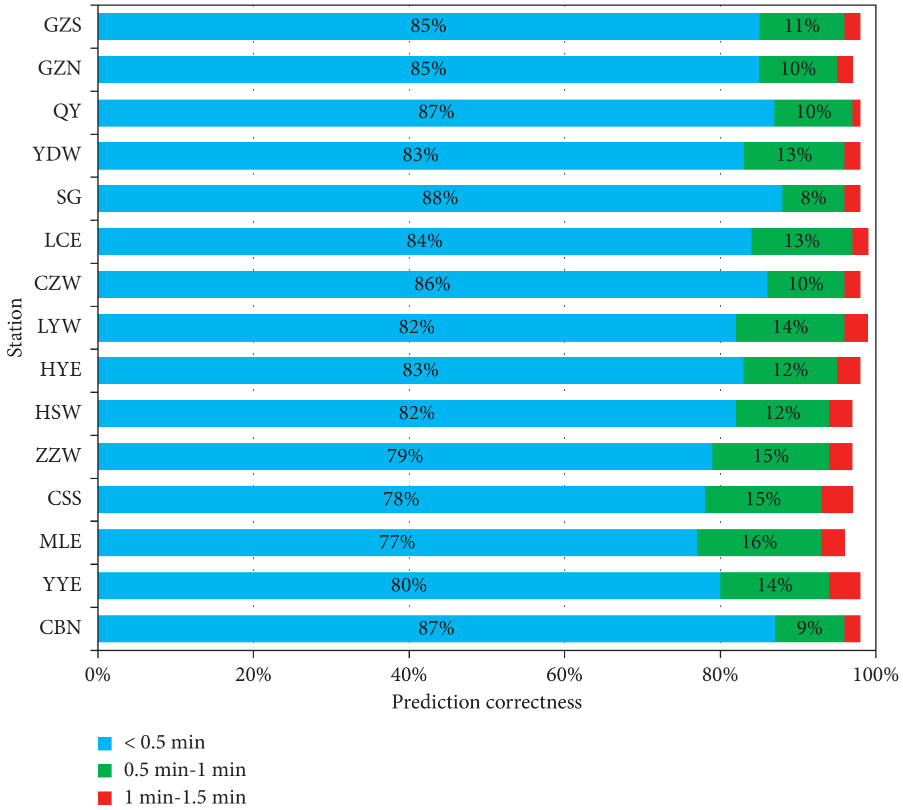
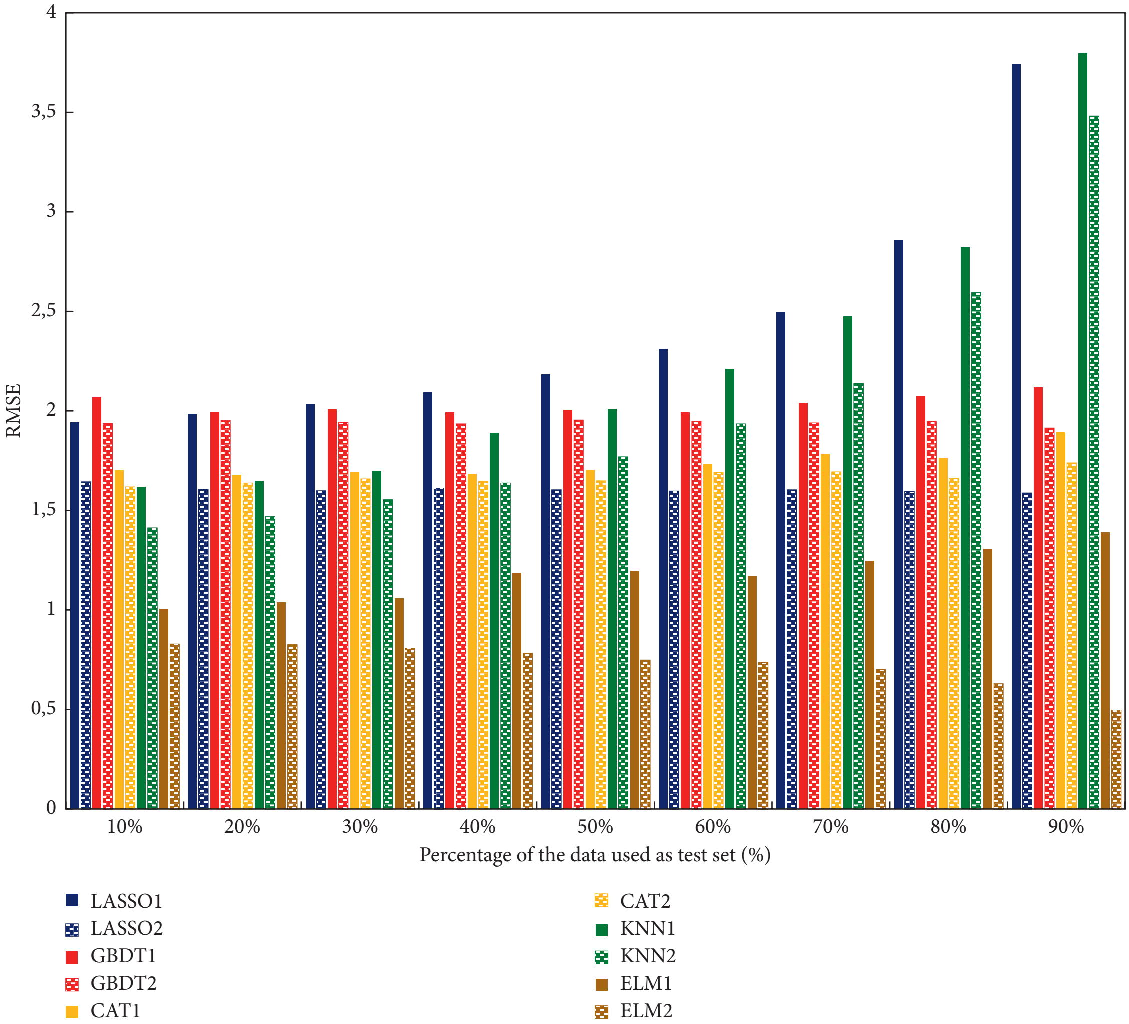
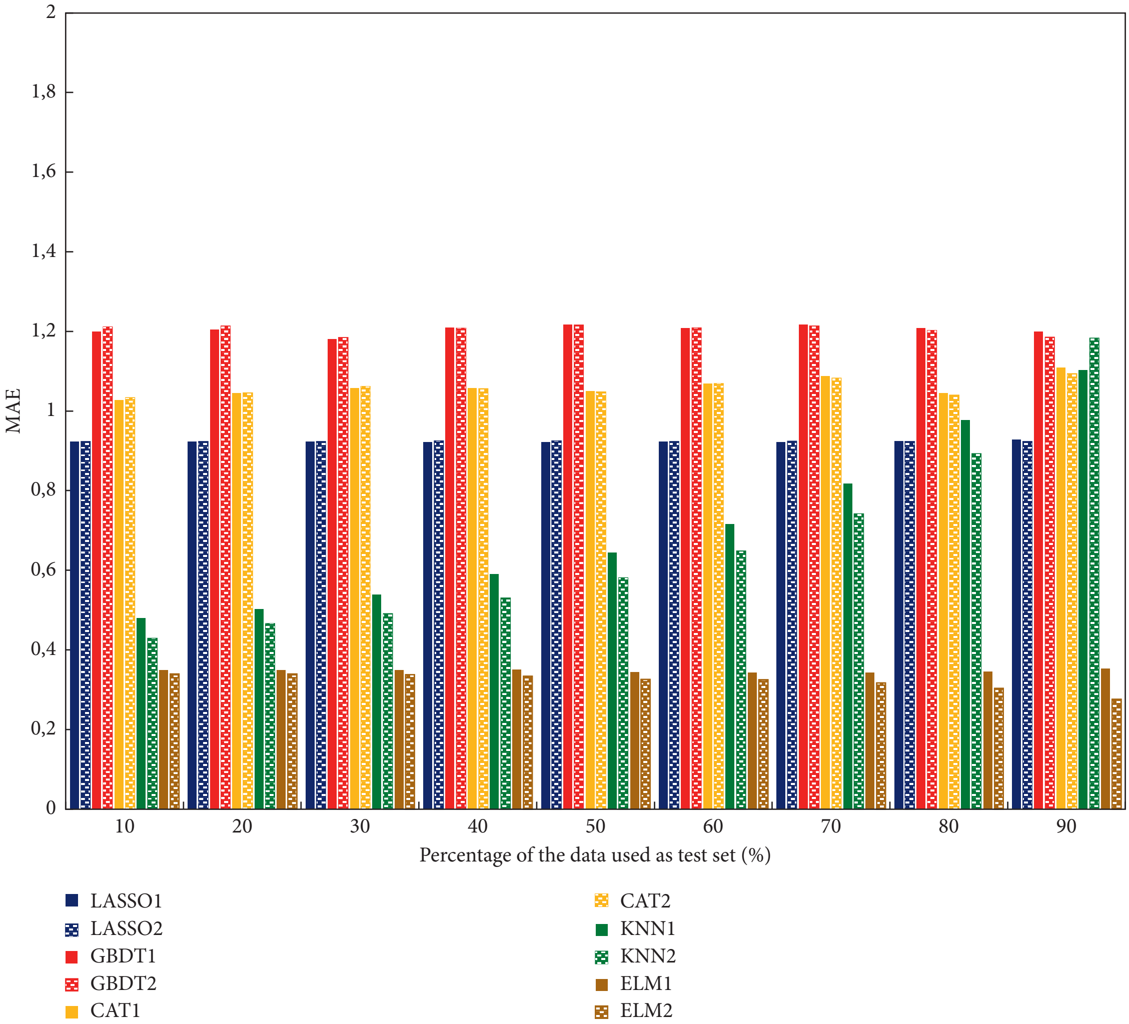


Figure 8: Prediction correctness for each station on test set.



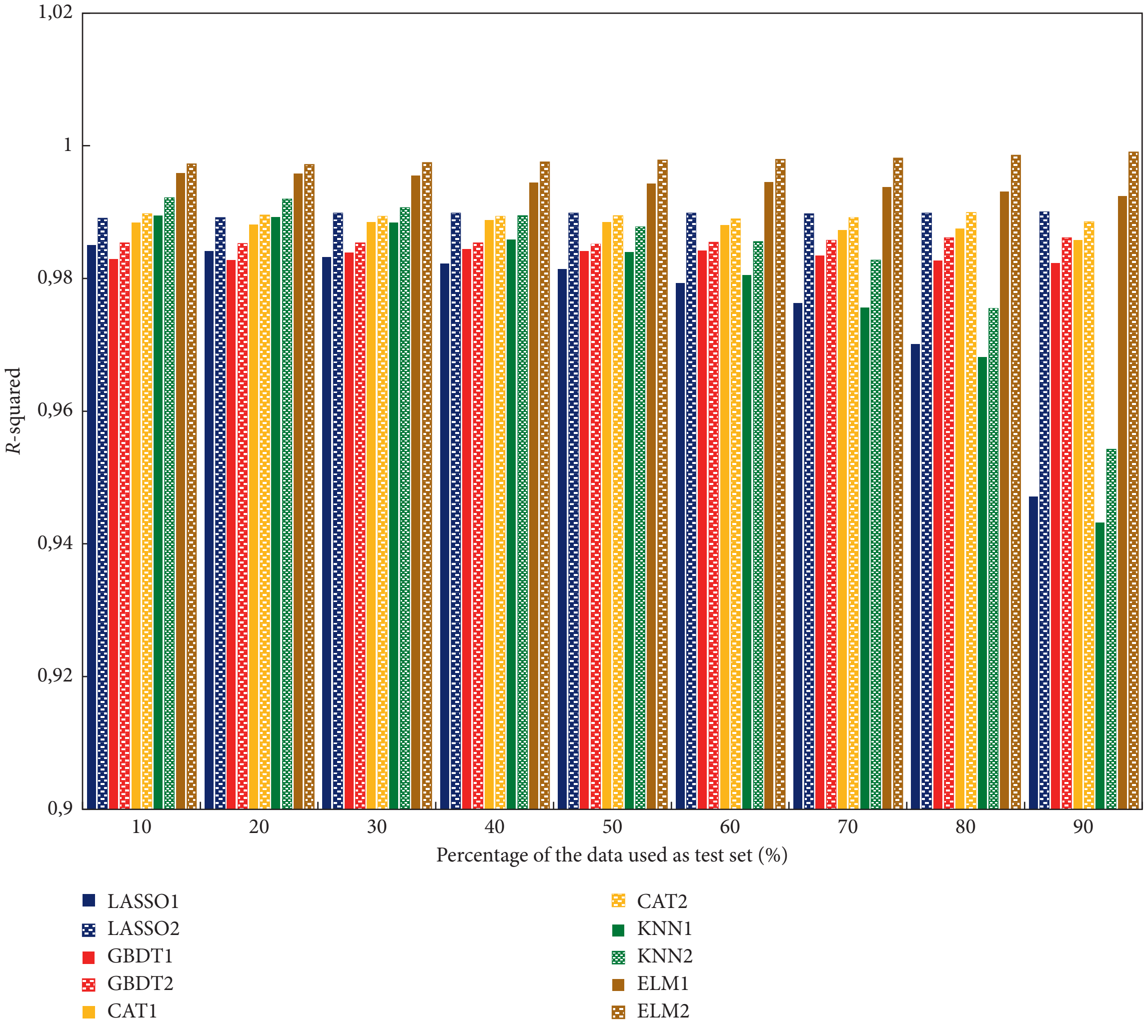
(a)

Figure 9: Continued.



(b)

Figure 9: Continued.



(c)

Figure 9: MAE, RMSE, and *R*-squared values on training and test sets with different data sizes (LASSO1 represents the performance on training set; LASSO2 represents the performance).

Table 6: Friedman ranking test and WSRT results.

Models	Mean rank	FT p value	Model 1 vs. model 2	WSRT Z-score	WSRT p value
ELM-PSO (M1)	1.50		—	—	—
CB (M2)	3.55		M1-M2	−259.200	≤0.001
GBDT (M3)	4.08	≤0.001	M1-M3	−261.312	≤0.001
Lasso (M4)	3.53		M1-M4	−257.456	≤0.001
KNN (M5)	2.34		M1-M5	−174.373	≤0.001

6. Conclusion

In this paper, a hybrid ELM-PSO method is proposed to predict train delays. The ELM can overcome the shortcomings of backpropagation training algorithm, and the advantage of PSO is its excellent ability in searching the best hyperparameter. Four benchmark models, CB, KNN, GBDT, and Lasso models, are selected to compare with proposed model. These models were run on the same data collected from China Railways. ELM-PSO tends to have a better performance and generalization ability (*R*-squared = 0.9955, MAE = 0.3490, RMSE = 1.0387) than the other models on the test set. Our work can not only provide sufficient time and auxiliary decision for the dispatcher to make reasonable optimization and adjustment plan, but also have practical significance for improving the quality of railway service and helping passengers estimate their travel time.

The dataset used in this paper contains train delays under all types of scenarios. Therefore, in the future, we will consider dividing all the data into certain types of delay scenarios according to particular rules and implementing currently prevalent models to train and predict each scenario to achieve a higher accuracy. Finally, in terms of the input features, all the information of the features in this paper can be obtained from train timetables. In the future, other types of features, such as the infrastructure, weather features, and other HSR lines obstruction, will be taken into account.

Data Availability

The data used to support the findings of this study were supplied by China Railway Guangzhou Bureau Group Co. Ltd. under license and so cannot be made freely available. Access to these data should be considered by the corresponding author upon request, with permission of China Railway Guangzhou Bureau Group Co. Ltd.

Conflicts of Interest

The authors declare that they do not have any commercial or associative interest that represents conflicts of interest in connection with the paper they submitted.

Authors' Contributions

Xu Bao contributed to conceptualization, prepared the original draft, was responsible for software, and visualized the study. Yanqiu Li prepared the original draft, was responsible for software, and visualized the study. Jianmin Li contributed to methodology and reviewed and edited the manuscript. Rui Shi contributed to supervision and data curation. Xin Ding contributed to data curation.

Acknowledgments

This work was financially supported by the Fundamental Research Funds for the Central Universities of China (2019JBM077) and the Open Fund for Jiangsu Key Laboratory of Traffic and Transportation Security (Huaiyin Institute of Technology).

References

[1] J. L. Espinosa-Aranda and R. García-Ródenas, "A demand-based weighted train delay approach for rescheduling railway networks in real time," *Journal of Rail Transport Planning & Management*, vol. 3, no. 1-2, pp. 1–13, 2013.

[2] N. Marković, S. Milinković, K. S. Tikhonov, and P. Schonfeld, "Analyzing passenger train arrival delays with support vector regression," *Transportation Research Part C: Emerging Technologies*, vol. 56, pp. 251–262, 2015.

[3] C. Wen, W. Mou, P. Huang, and Z. Li, "A predictive model of train delays on a railway line," *Journal of Forecasting*, vol. 39, no. 3, pp. 470–488, 2019.

[4] A. Higgins and E. Kozan, "Modeling train delays in urban networks," *Transportation Science*, vol. 32, no. 4, pp. 346–357, 1998.

[5] J. Yuan and I. A. Hansen, "Optimizing capacity utilization of stations by estimating knock-on train delays," *Transportation Research Part B: Methodological*, vol. 41, no. 2, pp. 202–217, 2007.

[6] P. B. L. Wiggenraad, *Alighting and Boarding Times of Passengers at Dutch Railway Stations*, TRAIL Research School, Delft, Netherland, 2001.

[7] M. F. Gorman, "Statistical estimation of railroad congestion delay," *Transportation Research Part E: Logistics and Transportation Review*, vol. 45, no. 3, pp. 446–456, 2009.

[8] P. Huang, C. Wen, and L. Fu, "Modeling train operation as sequences: a study of delay prediction with operation and weather data," *Transportation Research Part E: Logistics and Transportation Review*, vol. 141, 2020.

[9] P. Kecman and R. M. P. Goverde, "Online data-driven adaptive prediction of train event times," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 1, pp. 465–474, 2015.

[10] S. Milinković, M. Marković, S. Vesković et al., "A fuzzy Petri net model to estimate train delays," *Simulation Modelling Practice & Theory*, vol. 33, pp. 144–157, 2013.

[11] J. Peters, B. Emig, M. Jung, and S. Schmidt, "Prediction of delays in public transportation using neural networks," in *Proceedings of the International Conference on Computational Intelligence for Modelling, Control and Automation and International Conference on Intelligent Agents, Web Technologies and Internet Commerce*, vol. 2, pp. 92–97, Vienna, Austria, November 2005.

[12] C. Jiang, P. Huang, J. Lessan, L. Fu, and C. Wen, "Forecasting primary delay recovery of high-speed railway using multiple linear regression, supporting vector machine, artificial neural network, and random forest regression," *Canadian Journal of Civil Engineering*, vol. 46, no. 5, pp. 353–363, 2019.

[13] J. Hu and B. Noche, "Application of artificial neuron network in analysis of railway delays," *Open Journal of Social Sciences*, vol. 4, no. 11, pp. 59–68, 2016.

[14] M. Yaghini, M. M. Khoshraftar, and M. Seyedabadi, "Railway passenger train delay prediction via neural network model," *Journal of Advanced Transportation*, vol. 47, no. 3, pp. 355–368, 2013.

[15] F. Corman and P. Kecman, "Stochastic prediction of train delays in real-time using Bayesian networks," *Transportation Research Part C: Emerging Technologies*, vol. 95, pp. 599–615, 2018.

[16] J. Lessan, L. Fu, and C. Wen, "A hybrid Bayesian network model for predicting delays in train operations," *Computers & Industrial Engineering*, vol. 127, pp. 1214–1222, 2019.

[17] P. Huang, C. Wen, L. Fu, Q. Peng, and Y. Tang, "A deep learning approach for multi-attribute data: a study of train delay prediction in railway systems," *Information Sciences*, vol. 516, pp. 234–253, 2020.

[18] G. Bin Huang, Q. Y. Zhu, and C. K. Siew, "Extreme learning machine: theory and applications," *Neurocomputing*, vol. 70, no. 1–3, pp. 489–501, 2006.

[19] Y. Chen and W. Wu, "Mapping mineral prospectivity using an extreme learning machine regression," *Ore Geology Reviews*, vol. 80, pp. 200–213, 2017.

[20] Y. Yoan Miche, A. Sorjamaa, P. Bas et al., "OP-ELM: optimally pruned extreme learning machine," *IEEE Transactions on Neural Networks*, vol. 21, no. 1, pp. 158–162, 2010.

[21] Y. Yimin Yang, Y. Yaonan Wang, and X. Xiaofang Yuan, "Bidirectional extreme learning machine for regression problem and its learning effectiveness," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 23, no. 9, pp. 1498–1505, 2012.

[22] F. Han, H.-F. Yao, and Q.-H. Ling, "An improved evolutionary extreme learning machine based on particle swarm optimization," *Neurocomputing*, vol. 116, pp. 87–93, 2013.

[23] L. Oneto, E. Fumeo, G. Clerico et al., "Dynamic delay predictions for large-scale railway networks: deep and shallow extreme learning machines tuned via thresholdout," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 47, no. 10, pp. 2754–2767, 2017.

[24] Y. Li and Z. Yang, "Application of EOS-ELM with binary jaya-based feature selection to real-time transient stability assessment using PMU data," *IEEE Access*, vol. 5, pp. 23092–23101, 2017.

[25] Y. Zhang, T. Li, G. Na, G. Li, and Y. Li, "Optimized extreme learning machine for power system transient stability prediction using synchrophasors," *Mathematical Problems in Engineering*, vol. 2015, Article ID 529724, 8 pages, 2015.

[26] Y. Wang, H. Zhang, and G. Zhang, "cPSO-CNN: an efficient PSO-based algorithm for fine-tuning hyper-parameters of convolutional neural networks," *Swarm and Evolutionary Computation*, vol. 49, pp. 114–123, 2019.

[27] S. V. Konstantinov, A. I. Diveev, G. I. Balandina, and A. A. Baryshnikov, "Comparative research of random search algorithms and evolutionary algorithms for the optimal control problem of the mobile robot," *Procedia Computer Science*, vol. 150, pp. 462–470, 2019.

[28] D. Wang, D. Tan, and L. Liu, "Particle swarm optimization algorithm: an overview," *Soft Computing*, vol. 22, no. 2, pp. 387–408, 2017.

[29] Y. Li, X. Xu, J. Li, and R. Shi, "A delay prediction model for high-speed railway: an extreme learning machine tuned via particle swarm optimization," in *Proceedings of the 2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*, Rhodes, Greece, September 2020.

[30] M. Perceptron, J. Tang, and S. Member, "Extreme learning machine for multilayer perceptron," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 27, no. 4, pp. 809–821, 2016.

[31] J.-R. Zhang, J. Zhang, T.-M. Lok, and M. R. Lyu, "A hybrid particle swarm optimization-back-propagation algorithm for feed forward neural network training," *Applied Mathematics and Computation*, vol. 185, no. 2, pp. 1026–1037, 2007.

[32] Z. Deng, X. Zhu, D. Cheng, M. Zong, and S. Zhang, "Efficient kNN classification algorithm for big data," *Neurocomputing*, vol. 195, pp. 143–148, 2016.

[33] Y. Zhang, Z. Zhao, and J. Zheng, "CatBoost: a new approach for estimating daily reference crop evapotranspiration in arid and semi-arid regions of Northern China," *Journal of Hydrology*, vol. 588, Article ID 125087, 2020.

[34] G. Huang, L. Wu, X. Ma et al., "Evaluation of CatBoost method for prediction of reference evapotranspiration in humid regions," *Journal of Hydrology*, vol. 574, pp. 1029–1041, 2019.

[35] J. Yang, C. Zhao, H. Yu, and H. Chen, "Use GBDT to predict the stock market," *Procedia Computer Science*, vol. 174, pp. 161–171, 2020.

[36] R. Tibshirani, "Regression shrinkage and selection via the lasso: a retrospective," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 73, no. 3, pp. 273–282, 2011.

[37] M. Wang, L. Wang, X. Xu, Y. Qin, and L. Qin, "Genetic algorithm-based particle swarm optimization approach to reschedule high-speed railway timetables: a case study in China," *Journal of Advanced Transportation*, vol. 2019, Article ID 6090742, 12 pages, 2019.

[38] R. Shi, X. Xu, J. Li et al., "Prediction and analysis of train arrival delay based on XGBoost and Bayesian optimization," *Applied Soft Computing*, vol. 109, 2021.

Hindawi
Journal of Advanced Transportation
Volume 2021, Article ID 5574093, 21 pages
https://doi.org/10.1155/2021/5574093

Research Article

Data-Driven Approach for Passenger Mobility Pattern Recognition Using Spatiotemporal Embedding

Chao Yu,[1,2] **Haiying Li**,[1] **Xinyue Xu**,[1] **Jun Liu**,[1] **Jianrui Miao**,[1] **Yitang Wang**,[3] **and Qi Sun**[4]

[1]*State Key Laboratory of Rail Traffic Control & Safety, Beijing Jiaotong University, Beijing 100044, China*
[2]*School of Traffic and Transportation, Beijing Jiaotong University, Beijing 100044, China*
[3]*Track and Transportation Department, China Railway First Survey and Design Institute, Xi'an, Shaanxi 710043, China*
[4]*ACC Technical Room, Beijing Metro Network Control Center, Beijing 100101, China*

Correspondence should be addressed to Xinyue Xu; xxy@bjtu.edu.cn

Received 8 January 2021; Revised 11 April 2021; Accepted 8 May 2021; Published 19 May 2021

Academic Editor: Chung-Cheng Lu

Urban mobility pattern recognition has great potential in revealing human travel mechanism, discovering passenger travel purpose, and predicting and managing traffic demand. This paper aims to propose a data-driven method to identify metro passenger mobility patterns based on Automatic Fare Collection (AFC) data and geo-based data. First, Point of Information (POI) data within 500 meters of the metro stations are captured to characterize the spatial attributes of the stations. Especially, a fusion method of multisource geo-based data is proposed to convert raw POI data into weighted POI data considering service capabilities. Second, an unsupervised learning framework based on stacked auto-encoder (SAE) is designed to embed the spatiotemporal information of trips into low-dimensional dense trip vectors. In detail, the embedded spatiotemporal information includes spatial features (POI categories around the origin station and that around the destination station) and temporal features (start time, day of the week, and travel time). Third, a density-based clustering algorithm is introduced to identify passenger mobility patterns based on the embedded dense trip vectors. Finally, a case of Beijing metro network is used to verify the feasibility of the above methodology. The results show that the proposed method performs well in recognizing mobility patterns and outperforms the existing methods.

1. Introduction

The number of urban residents is increasing significantly, and human mobility is becoming unpredictable and complex, posing major challenges to public safety and health (such as the COVID-19 epidemic). In recent years, urban mobility pattern recognition has become a hotspot due to its ability to reveal resident life routines, assist in transportation planning, estimate and manage travel demand, predict passenger travel purposes, and provide location-based services [1–5]. As an important part of urban transportation, the metro system has increasingly become an indispensable choice for urban residents. Therefore, studying metro passenger mobility patterns is essential for analyzing urban mobility characteristics.

Fortunately, the continuous development of digitalization has provided strong support for urban planning and transportation services. Currently, large-scale spatiotemporal travel-related data provide the possibility for the analysis of passenger mobility patterns. From the perspective of the types of raw data, the recognition of urban mobility patterns can be divided into two categories, namely, researches based on trajectory data and that based on AFC data. The former is mainly meant to reproduce the movement track of residents through GPS data, social media data, or mobile phone signaling data to identify mobility patterns [6–12]. Unlike this, the latter often uses the tap-in or tap-out data of passengers to describe the travel process in order to realize the analysis of travel patterns [1, 13–20]. However, there are some shortcomings in trajectory data. First, trajectory data are often obtained when the mobile phone user turns on the positioning function, which means that the behavior of the user to turn on or off the positioning function has a direct impact on the collection of trajectory data. Second, the accuracy of trajectory data depends on the reliability of positioning technology. In fact, most positioning methods often have unavoidable errors, especially in densely populated areas or underground multistory buildings, resulting in blurred trajectories. Conversely, an individual trajectory identified by AFC data is error-free at the spatial level of stops and stations [15]. Admittedly, AFC data cannot pinpoint the specific activity location of passengers. However, it is possible to use the land-use data around the station to infer the possible activity locations of passengers, because passengers often complete the displacement before tap-in or after tap-out by walking [21].

It is undeniable that trajectory data and AFC data have their own advantages and disadvantages in identifying passenger mobility patterns. For metro managers and operators, AFC data are relatively accurate and easily available. Using AFC data to analyze passenger mobility patterns and behavior characteristics can significantly improve the metro service level. This paper aims to propose a data-driven approach to explore the possibility of AFC data in inferring passenger mobility patterns. In the existing research on mobility patterns, the tap-in timestamp, tap-out timestamp, and travel time are usually fused to mine the temporal characteristics. However, the discovery of spatial features usually stays at the station level. The common method is to characterize the latent spatial characteristics by dividing the stations into several different clusters, which makes it difficult to infer the specific mobility patterns of passengers. In view of the above analysis, AFC data are selected to extract passenger travel information. In addition, multisource geo-based data are captured to provide the necessary land-use information to realize passenger mobility recognition. In this paper, each AFC travel record is processed by an unsupervised method into a low-dimensional vector containing spatiotemporal features. There are two advantages. First, the concrete spatial information and temporal information being transformed into abstract vector forms are convenient for large-scale processing by computers (for example, similarity calculation). Second, vectorization can extract the characteristics of travel records to the maximum extent while saving storage space to explore the internal mechanism of passenger mobility [7].

The contribution of this paper is threefold. First, a multisource data fusion method is presented. This method adds the residential area information provided by the housing trading platform and the building information provided by the geographic information service to the raw POI data to convert the raw POI data into weighted POI data considering service capabilities. It avoids the drawbacks of using POI numbers to quantify land-use characteristics in the existing works [21]. Second, an unsupervised deep learning framework based on SAE is proposed to embed the spatiotemporal information of passenger travel, so as to realize the conversion of a passenger travel record into a low-dimensional dense vector. In this framework, the self-encoding is utilized to realize the embedding of spatiotemporal information without the labeled data and supervised training, which can extract the features of travel records more comprehensively than existing methods [22, 23]. Third, a density-based clustering algorithm is used to identify passenger mobility patterns. It can generate the number of clusters according to the data distribution without manually specifying the number of clusters, avoiding the human intervention of existing methods [7, 24].

The structure of this paper is as follows. In Section 2, the existing studies on mobility pattern recognition are classified and summarized. In Section 3, the methodology of this paper is introduced in detail, including an overview of the method and three main steps, namely, the fusion of multisource geo-based data, embedding spatiotemporal semantics in trip records, and mobility pattern recognition based on the embedded vectors. In Section 4, a case based on the Beijing metro network is introduced to verify the effectiveness of the proposed method, and the results of the case study are compared with existing methods. Besides, potential applications based on passenger mobility pattern recognition are explained. Finally, the paper is summarized and discussed in Section 5.

2. Literature Review

Passenger mobility pattern recognition aims to discover the identifiable travel categories formed by passengers in the long-term travel history, such as working, going home, entertainment, etc. Existing research has revealed that urban mobility exhibits a high degree of regularity in time and space [7, 25]. This allows us to discover the daily routines and social state of travelers through mobility analysis. To do this, many methods have been proposed in the existing work. Macroscopically, these methods can be classified into two categories, namely, empirical models and data-driven models.

Intuitively, the empirical method is to quantitatively analyze passenger behavior by features or thresholds of the known activity categories. The abovementioned features and thresholds tend to be artificially designated by researchers or experts. For example, a rule was established by [18] that the cardholder's first tap-in station or the last tap-out station can be considered as his/her potential home location. An algorithm based on "center point" is proposed to infer cardholder's exact home location based on multiple potential locations. The effectiveness of this method is verified by a case of Beijing metro system, in which 88.7% of passengers' home locations were successfully inferred. Similarly, a passenger's home location was determined to be the most visited location between 7 pm and 8 am on weekends and weekdays, as suggested by [11]. It was presented by [9] that a passenger's home and work place are the most visited and second most visited locations. Although the above assumptions can help infer the passenger's home and work locations to a certain extent, they are not universal. The rules are often subjective, and their application effects rely heavily on the domain knowledge of experts or scholars [23]. Furthermore, the empirical method is incapable of discovering new mobility patterns, resulting in the inability to keenly estimate the changing trend of urban mobility with the increase in population and the complexity of the urban transportation network.

In order to avoid the above shortcomings, data-driven methods have emerged. As mentioned in Section 1, large-scale datasets provided more possibilities for mobility analysis. In the past few years, a variety of datasets have been used to describe urban mobility, such as mobile phone signing data, GPS data, media data, AFC data, sociodemographic data, and census and administrative data [1, 26, 27]. Faced with such diverse datasets, many data-driven methods have also been proposed by researchers to mine passenger mobility patterns. For instance, multi-objective Convolutional Neural Network (CNN) was designed to infer the social demographic attributes and mobility features of passengers based on media data and land-use data [28]. Support vector machines (SVM) were introduced to divide passenger travel data into several types, and the passenger purpose was analyzed according to the characteristics of each type using sociodemographic data [8]. This method was applied to data from a large number of Californian families. The application results showed that this method performed better than the traditional multinomial logit models. Moreover, smart card data can also be utilized to construct land-use function complementation indices to improve the performance of the classic gravity model in analyzing the human mobility between different types of areas in the city [29]. The case of Shenzhen metro showed that these indices were effective tools to reveal the mechanism of spatial interaction and had a significant effect on improving the prediction of spatial flow and travel distribution. The naive Bayes probability model was improved to observe the continuous long-term changes in the attributes of metro passenger trips using AFC data and census data [13]. The verification results of real cases showed that 86.2% of passengers' travel purpose can be estimated. A data-driven robust method using AFC data and the General Transportation Feedback Specification (GTFS) was designed to infer the most likely movement trajectory of each passenger [20]. The use of GTFS data reduced many assumptions about the passenger travel process in previous studies (the threshold assumptions of transfer travel time, time window assumptions for selecting vehicles and journeys, threshold assumptions for waiting and boarding time, etc.). This method was used in the analysis of passenger travel trajectories in Minnesota and proved to be superior to traditional trajectory inference methods. Besides, to recognize the patterns of passengers' variation over time and analyze the spatial heterogeneity of the dynamic space around the metro stations, an eigendecomposition method was proposed [30]. In this work, the datasets were decomposed into a combination of principal components and eigenvectors, where the principal components represent the common pattern of passenger movement, and the corresponding elements in the eigenvectors mean the attributes of metro stations. The above method was verified in the case of the Shenzhen metro system and proved to be effective in improving urban planning. A method based on the Hidden Markov Model (HMM) was addressed to infer the sequence of passenger activities, and the model parameters were calibrated using Baum–Welch algorithm based on land-use data around the stations [31]. The abovementioned data-driven methods excavated the rules of passenger mobility from different aspects, but there are still shortcomings of high computational cost and poor interpretability.

In recent years, various types of topic models have gradually become the mainstream methods for the analysis of urban mobility patterns [6, 9, 23]. In these studies, mobility pattern recognition was regarded as a topic mining problem in the field of natural language processing (NLP). In the model, each passenger was treated as an article, each trip record of the passenger is processed as a word in the article, and the previous and subsequent trips of a certain trip were considered as the context of the current trip. Correspondingly, passenger mobility pattern recognition can be understood as mining several topics in the corpus composed of multiple articles. For example, a multi-directional probabilistic factorization model based on tensor decomposition and probabilistic latent semantic analysis (PLSA) was proposed, which used a simple latent semantic structure to describe the multi-directional mobility characteristics of passengers involved in high-order interactions [16]. The multi-directional mobility analysis of urban residents in Singapore verified the practicality of the model. A Bayesian n-gram model was constructed to predict the location and time of individual passenger activities, and its prediction result was expressed as an ordered set of passenger potential activities, which contains the location and time of each activity [32]. On this basis, a spatiotemporal topic model based on Latent Dirichlet Allocation (LDA) was presented to classify passenger activities into several topics to realize mobility pattern recognition [23]. The above method was verified by the travel data of more than 10,000 users of the London Underground in 2 years, and the results showed that the median accuracy of travel prediction could reach 80%. The obtained passenger mobility patterns could well reveal the temporal and spatial attributes of work-related and home-related activities. Unfortunately, the abovementioned researches only analyzed mobility from the perspective of temporal characteristics, without considering spatial information, which makes the results poor in interpretability. Considering spatial features, methods based on word vector were introduced for exploring mobility patterns. For example, a habit2vec method was proposed by [7] to embed a passenger's current visit to a POI type during a time slice. Besides, the inbound flow, the outbound flow, and the surrounding POIs were used as elements to construct the target station vector suggested by [21]. In this work, it was worth noting that the Term Frequency–Inverse Document Frequency (TF-IDF), which was an indicator in the NLP field, was applied to quantify categories of the target station. Nevertheless, it is unreasonable to determine station categories only by the frequency or TF-IDF of different categories of POI around the station due to the significant difference in service capabilities of different categories of POI. For example, although a residential area and a cafe are both displayed as POIs on the map, the service capacity of the former is obviously greater than that of the latter. Therefore, a POI needs to be weighted according to its service capability to be meaningful in describing passenger mobility.

In a nutshell, the existing works on passenger mobility is in the ascendant, but there are still defects such as high computational cost, lack of consideration of spatial features, and poor interpretability. In this paper, weighted POI is first generated through multisource geo-based data. Then, through the unsupervised learning framework based on SAE, the temporal and spatial features are simultaneously embedded into the trip vector of passengers to identify the mobility patterns. The following is the methodology of this work.

3. Methodology

The overview of the methodology is shown in Figure 1. The goal is to design an efficient method to transform trip records into standard forms that can be processed by computers, so as to simplify mobility pattern recognition into a clustering problem. After obtaining AFC records, the following three steps are required to achieve the above goal. First, a fusion of multisource, geo-based data method is proposed to weight the raw POI data and provide a basis for spatial semantic estimation. Second, a low-dimensional dense trip vector containing both spatial and temporal attributes is generated to represent the given record. Third, clustering analysis on low-dimensional dense trip vectors is addressed to distinguish between different trip clusters to realize mobility pattern recognition. Details of these three steps are described in the following sections.

3.1. Fusion of Multisource Geo-Based Data. POI is a point unit in geographic information systems to mark the location of human activity. A POI contains the POI name, category label, longitude, latitude, and land-use type information of the point unit [1]. Some existing studies infer the travel purpose of passengers through the category label of POIs around the target station. For example, when the POIs around a passenger's origin station are mostly residential and the POIs around the destination station are mostly working, the passenger's travel purpose can be considered to have a high probability of going to work [21]. Note that a POI can be a residential neighborhood, a shopping center, or a kindergarten. The service capacity of a residential neighborhood is obviously greater than that of a kindergarten. So, it is inaccurate to infer travel purpose from the number of POIs due to the difference in service capacity of different types of POIs. The goal of this section is to generate weighted POIs considering service capacity using multisource, geo-based data.

The geo-based data involved in this paper are obtained from three data sources, namely, Amap, Lianjia, and Arctiler. Among them, Amap (https://www.amap.com/) is a provider of digital map content, navigation, and location services solutions. It provides the raw POI data. It should be noted that Amap divides all POIs into 24 categories. For details of the classification, please refer to the website (https://lbs.amap.com/api/webservice/download). In this paper, from the perspective of travel purpose, these categories are integrated into 8 categories, as shown in Table 1. In addition, some POIs that are not closely related to the travel purpose, such as public toilet and traffic light, are deleted. Besides, Lianjia (https://www.lianjia.com/) is a housing trading platform that can provide the neighborhood properties containing the name, housing price, property management fee, the number of buildings, and the number of households in a targeted residential neighborhood. For the residential POI in Table 1 (category 6), we use the number of households to represent its actual service capacity. Further, Arctiler (http://www.arctiler.com/) is a geographic information service provider that can provide the building physical properties containing the name, building category, usable area, and the number of floors of a target building. For different types of buildings, the per capita service area is stipulated by the Technical Measures for National Civil Building Engineering Design (http://www.chinabuilding.com.cn/book-1815.html). Therefore, we can calibrate the actual service capacity of the nonresidential POI in Table 1 by combining the building physical properties and per capita service area. With the above processing, the raw POI data have been converted into weighted POI data considering service capacity.

It should be noted that due to different data sources, the POI name may be different from the building name or the residential area name for the same point unit on the map, making data fusion difficult to achieve. Here, a data matching method is designed, as shown in Figure 2. For a given target POI, a building is selected from the Arctiler database, and the distance between the two is calculated to determine whether it matches each other. Note that it is necessary to convert the longitude and latitude of the building base outline obtained from Arctiler to that of the building base center. And then, the actual distance between the two coordinates can be calculated as follows:

$$\text{distance}\,(A, B) = \theta_{A,B} \cdot \frac{2\pi}{360} \cdot R_{\text{Earth}} \cdot 1000, \tag{1}$$

$$\begin{aligned}\theta_{A,B} = {} & \arccos\left(\cos\left(A.\text{lat}\right)\cos\left(B.\text{lat}\right)\cos\left(A.\ \ln g\right.\right.\\ & \left.\left.-B.\ \ln g\right) + \sin\left(A.\text{lat}\right)\sin\left(B.\text{lat}\right)\right),\end{aligned} \tag{2}$$

where Distance (A, B) represents the actual distance between the two coordinate points A and B, in meters, $A.\text{lat}\,(B.\text{lat})$ and $A.\text{lng}\,(B.\text{lng})$ represent the latitude and longitude of A (B), and R_{Earth} represents the radius of the earth, which is 6371 km. All longitudes and latitudes in this paper are based on the World Geodetic System 1984 (WGS-84) coordinate system. Finally, it is judged whether the obtained distance is less than the threshold, which is set to 50 meters. If it is, the actual service capacity of the target POI is calibrated according to the per capita service area obtained from the Technical Measures for National Civil Building Engineering Design, that is, the weighted POI, otherwise, another building is selected from the Arctiler database to rematch the target POI. The data fusion process of residential POI is similar to this, and will not be repeated here. At this point, the raw POIs have been converted into weighted POIs based on multisource, geo-based data.

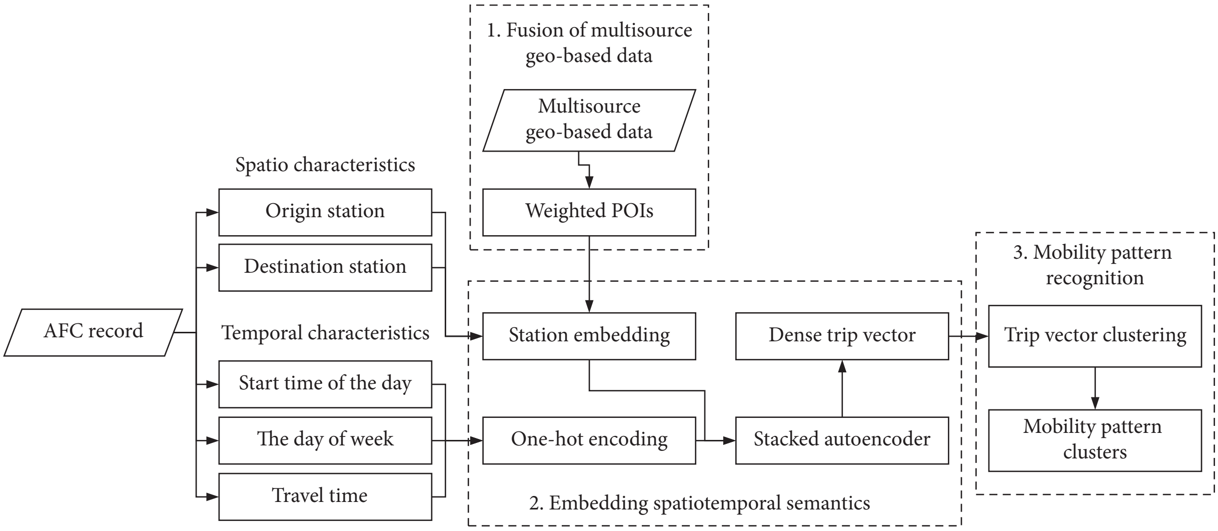


Figure 1: Overview of the methodology.

Table 1: POI categories and contents.

ID	Category	Contents
1	Entertainment	Recreation center, night club, KTV, disco, pub, game center, card and chess room, lottery center, Internet bar, recreation place, etc.
2	Working	Construction company, medical company, machinery and electronics, chemical and metallurgy, commercial trade, telecommunication company, mining company, etc.
3	Shopping	Shopping plaza, shopping center, shops, duty-free shop, convenience store, digital electronics, supermarket, plants and pet market, home building materials market, etc.
4	Transportation	Airport, railway station, passenger port, tourist routes bus station, common bus station, parking lot related, etc.
5	Education	Museum, exhibition Hall, convention and exhibition center, art gallery, library, planetarium, cultural palace, university and college, middle school, etc.
6	Residential	Hotel, residential area, villa, residential quarter, dormitory, community center, etc.
7	Hospital	Hospital, health center, clinic, disease prevention, pharmacy, medical supplies, etc.
8	Government	Governmental organization and institution, foreign embassy and consulate, representative office of international organizations, etc.

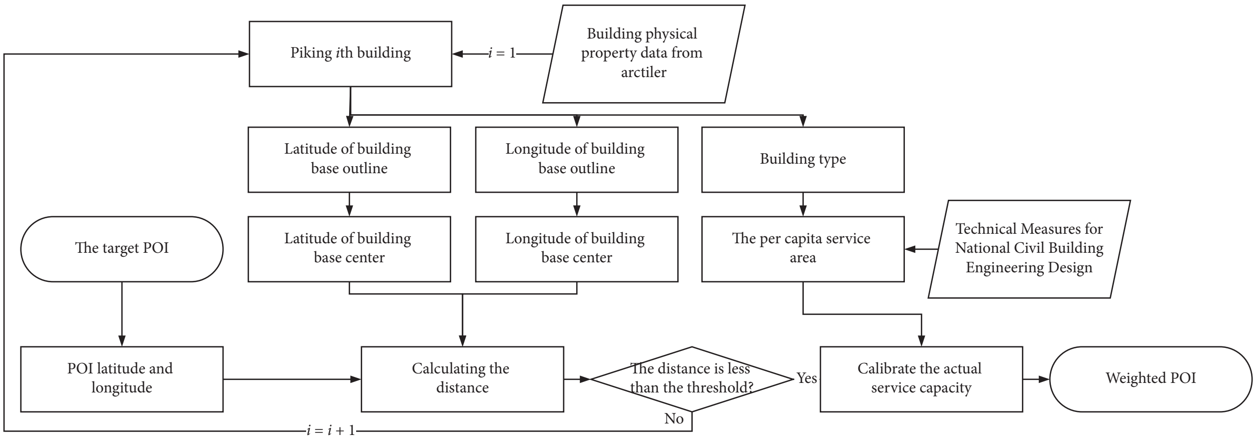

Piking ith building
i = 1
Building physical property data from arctiler
Latitude of building base outline
Longitude of building base outline
Building type
The target POI
Latitude of building base center
Longitude of building base center
The per capita service area
Technical Measures for National Civil Building Engineering Design
POI latitude and longitude
Calculating the distance
The distance is less than the threshold?
Yes
No
Calibrate the actual service capacity
Weighted POI
i = i + 1

Figure 2: Fusion of multisource, geo-based data.

3.2. Embedding Spatiotemporal Semantics in Trip Records. A passenger trip record R from AFC system is composed of four components, namely, the tap-in time t_{in}^{R}, the tap-in station s_{in}^{R}, the tap-out time t_{out}^{R}, and the tap-out station s_{out}^{R}. In this paper, the above four components are transformed into five attribute vectors to describe the passenger trip. They are the origin station vector O^{R}, the destination station $\mathbf{D}^{R}$, start time of the day $\mathbf{T}^{R}$, the day of week $\mathbf{W}^{R}$, and travel time $\mathbf{H}^{R}$. Symbolically, a trip record R corresponds to a vector $\mathbf{R}$, which can be represented as

$\{\mathrm{O}^R, \mathbf{D}^R, \mathbf{T}^R, \mathbf{W}^R, \mathbf{H}^R\}$. In this section, the goal is to represent the above attributes as spatiotemporal semantics in the form of vectors for subsequent mobility pattern recognition. To do this, a SAE-based framework is built to embed spatiotemporal semantics, the structure of which is shown in Figure 3. First, weighted POIs calibrated in Section 3.1 and one-hot encoding are addressed to generate spatial/temporal attribute vectors. Subsequently, the above vectors are assembled to form a high-dimensional sparse trip vector. This method proved to be reasonable and feasible [7, 21]. It should be noted that although the high-dimensional vector contains a variety of travel information, the sparsity makes the mobility pattern difficult to be recognized effectively. To solve this problem, we train a SAE model to transform the high-dimensional trip vector into a low-dimensional dense vector to represent spatiotemporal semantics. Here are the details.

In the existing researches, the radius of the service area of a metro station is generally set as 500 meters [18, 21]. Therefore, in terms of spatial semantic, weighted POIs within 500 meters of the target station are utilized to represent the station. Define P as the set of all weighted POIs in the research area. For the tap-in station s_{in}^R and the tap-out station s_{out}^R, the weighted POIs within 500 meters can be expressed as follows:

$$p_{s_{\mathrm{in}}^R} = \left\{p \middle| \mathrm{distance}\left(p, s_{\mathrm{in}}^R\right) \leq 500, \forall p \in P\right\}, \tag{3}$$

$$p_{s_{\mathrm{out}}^R} = \left\{p \middle| \mathrm{distance}\left(p, s_{\mathrm{out}}^R\right) \leq 500, \forall p \in P\right\}. \tag{4}$$

As shown in Table 1, the weighted POIs have been divided into 8 categories, so O^R and $\mathbf{D}^R$ can each be represented as an 8-dimensional vector. The value of a weighted POI represents its service capacity, and the larger the value, the greater the probability of becoming the departure point or destination point of passengers at the station. O^R and $\mathbf{D}^R$ can be expressed as follows:

$$\mathrm{O}^R = \left\{\frac{\sum|p_1|}{\sum\left|p_{\mathrm{in}}^R\right|}, \frac{\sum|p_2|}{\sum\left|p_{\mathrm{in}}^R\right|}, \ldots, \frac{\sum|p_8|}{\sum\left|p_{\mathrm{in}}^R\right|}\right\}, \quad p_i \in p_{\mathrm{in}}^R, i = 1, 2, \ldots, 8, \tag{5}$$

$$\mathbf{D}^R = \left\{\frac{\sum|p_1|}{\sum\left|p_{\mathrm{out}}^R\right|}, \frac{\sum|p_2|}{\sum\left|p_{\mathrm{out}}^R\right|}, \ldots, \frac{\sum|p_8|}{\sum\left|p_{\mathrm{out}}^R\right|}\right\}, \quad p_i \in p_{\mathrm{out}}^R, i = 1, 2, \ldots, 8. \tag{6}$$

where $\sum|p_i|$ represents the sum of value of weighted POI of the ith category, $\sum|p_{\mathrm{in}}^R|$ and $\sum|p_{\mathrm{out}}^R|$ represent the sum of all weighted POIs within 500 meters of the tap-in station s_{in}^R and the tap-out station s_{out}^R. The order of POI categories corresponds to the row order in Table 1, namely, Entertainment, Working, Shopping, Transportation, Education, Residential, Hospital, and Government.

As for temporal semantic, one-hot coding is adopted to represent three attributes. For the convenience of expression, we divide a day into several discrete slots with a fixed interval. The metro service is not available between 0 am and 5 am. Here, the interval is set to be one hour, resulting in 19 slots in a day. So $\mathbf{T}^R$ can be easily characterized as a 19-dimensional vector. For example, if t_{in}^R is $5:16:29$ (between 5 and 6), it can be expressed as $\{1, 0, 0, \ldots, 0\}$. If t_{in}^R is $22:51:33$ (between 22 and 23), it can be expressed as $\{0, 0, 0, \ldots, 0, 1, 0\}$. Similarly, because there are 7 days a week, $\mathbf{W}^R$ can be represented as a 7-dimensional vector. If t_{in}^R is on Monday, it can be expressed as $\{1, 0, 0, 0, 0, 0, 0\}$. As for travel time, since most passengers travel within 240 minutes, we divide the travel time into 8 slots with the interval of 30 minutes [33]. If the travel time of R is 57 minutes (between 30 and 60), $\mathbf{H}^R$ can be expressed as $\{0, 1, 0, 0, 0, 0, 0, 0\}$. In summary, the trip vector $\mathbf{R} = \{\mathrm{O}^R, \mathbf{D}^R, \mathbf{T}^R, \mathbf{W}^R, \mathbf{H}^R\}$ has been represented as a 50-dimensional $(8 + 8 + 19 + 7 + 8)$ sparse vector.

We train a SAE model to extract the mixed spatiotemporal semantics of trip record R to avoid the adverse effects of the sparsity of high-dimensional vector [34]. Essentially, the auto-encoder is an unsupervised algorithm that can automatically learn features from unlabeled data and can give a better feature description than the original data. It can be regarded as a neural network, which automatically generates an optimal coding strategy by continuously optimizing the weight parameters, resulting in the output vector being consistent with the input vector. As an extension of the classic auto-encoder, SAE is a deep neural network model constructed by stacking multiple auto-encoders, where the output of the nth layer of auto-encoder is used as the input of the $(n + 1)$th layer of auto-encoder [35]. Structurally, SAE can be divided into two components, namely, the encoder and decoder. The former transforms the input sparse vector into a dense vector through several layers of coding, and the latter is the reverse process of the former to reconstruct high-dimensional vectors. As shown in Figure 3, the input 50-dimensional sparse vector $\mathbf{R}$ is firstly upgraded to a 64-dimensional vector to extract abstract features, and then the dimensionality is reduced to 16-dimensional and 8-dimensional vectors layer by layer to realize the representation of dense vectors. Formulaically, the above process can be expressed as follows:

$$\mathbf{h}_{n+1} = f_a\left(\mathbf{W}_n \mathbf{h}_n + \mathbf{b}_n\right), \tag{7}$$

where $\mathbf{h}_n$ and $\mathbf{h}_{n+1}$ represent the output vector of the nth and the $(n + 1)$th layer, $\mathbf{W}_n$ and $\mathbf{b}_n$ represent the weight parameter

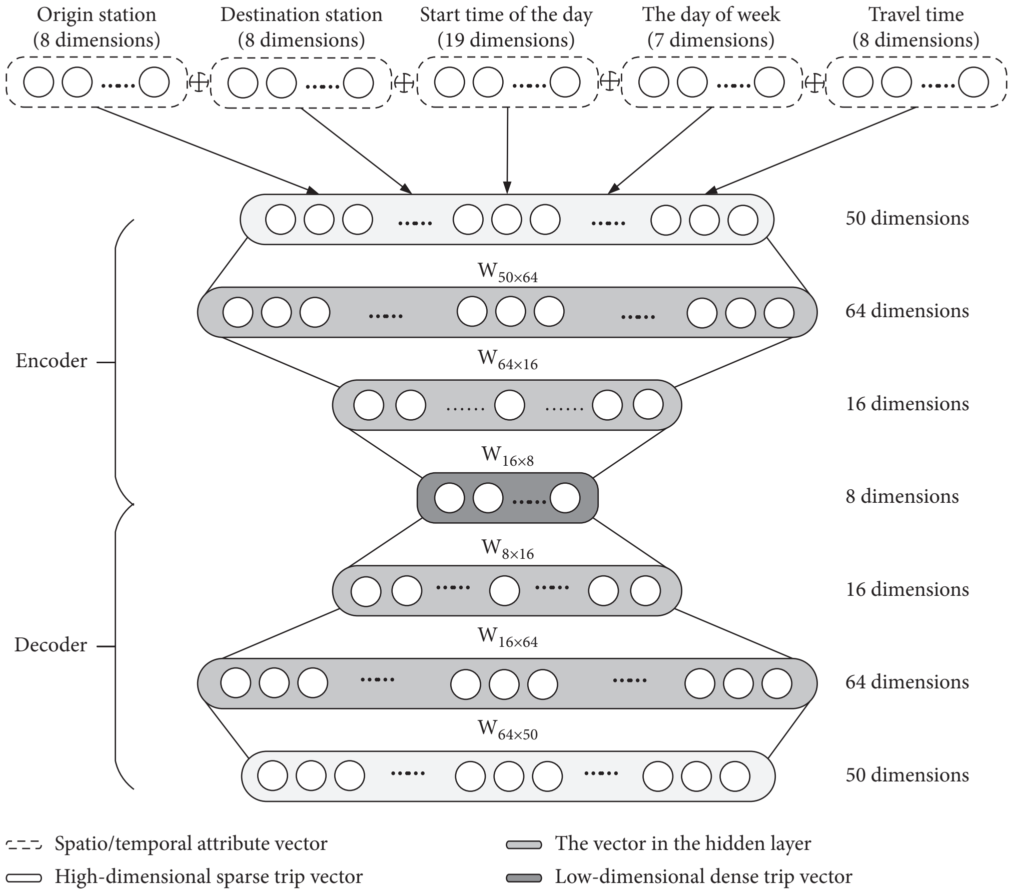

Origin station (8 dimensions)
Destination station (8 dimensions)
Start time of the day (19 dimensions)
The day of week (7 dimensions)
Travel time (8 dimensions)
50 dimensions
$W_{50\times64}$
64 dimensions
$W_{64\times16}$
16 dimensions
$W_{16\times8}$
8 dimensions
$W_{8\times16}$
16 dimensions
$W_{16\times64}$
64 dimensions
$W_{64\times50}$
50 dimensions
Encoder
Decoder
Spatio/temporal attribute vector
High-dimensional sparse trip vector
The vector in the hidden layer
Low-dimensional dense trip vector

Figure 3: Embedding spatiotemporal semantics using SAE.

matrix and the bias from the nth layer to the $(n+1)$th layer, and $f_a(\cdot)$ represents the activation function, which is the rectified linear unit (ReLU) in this paper. It can be seen that the parameters that need to be estimated in the model are $\mathbf{W}_n$ and $\mathbf{b}_n$. Particularly, when $n = 1$, $\mathbf{h}_n$ is $\mathbf{R}$. Since the dimension of h_1 is smaller than that of h_2 $(50 < 64)$, it is necessary to avoid invalid training of the weight parameters [36]. The weight parameters of this layer need to be pretrained, where the greedy layer-wise pre-training method is used. See details in reference [37]. The loss function is constructed as follows and the regularization is used in this process.

$$\text{loss} = L(x, g(f(x))) + \Omega(h). \tag{8}$$

Here, $h = f(x)$ represents the output of the encoder, and $g(f(x))$ represents the output of the decoder. Besides, $L(x, g(f(x)))$ represents the difference between x and $g(f(x))$, which can be measured by the mean square error (MSE). Further, $\Omega(h)$ represents the regularization term, which is the l_1-norm here. Using the above procedure, the weight parameters of this layer can be initialized. As for the weight parameters of other layers, truncated normal initializer can be used.

And then, MSE is chosen as the loss function of the whole SAE. Define the dense vector as $\mathbf{R}_{\text{dense}}$ and the output reconstruction high-dimensional vector as $\mathbf{R}_{\text{rc}}$, then the loss function f_{loss} can be expressed as follows:

$$f_{\text{loss}} = \text{MSE}(\mathbf{R}_{\text{rc}}) = \frac{\sum_N \left(\mathbf{R}_i - \mathbf{R}_{\text{rc},i}\right)^2}{N}, \tag{9}$$

where N represents the total number of trip records, $\mathbf{R}_i$ and $\mathbf{R}_{\text{rc},i}$ represent the ith element in vector $\mathbf{R}$ and $\mathbf{R}_{\text{rc}}$. As for training parameters, back propagation is used to fine-tune the parameters based on the value of the loss function. In this way, $\mathbf{R}$ is converted to $\mathbf{R}_{\text{dense}}$.

3.3. Mobility Pattern Recognition Based on the Embedded Vectors. The goal of this section is to cluster $\mathbf{R}_{\text{dense}}$ through the cluster algorithm and achieve mobility pattern recognition through the spatiotemporal characteristics (obtained by decoder) displayed by the clustering results. It is found that passenger trajectories tend to show a high degree of temporal and spatial regularity. Passengers follow simple reproducible patterns, indicating that each individual is characterized by a significant probability to return to a few highly frequented locations [38, 39]. Since the $\mathbf{R}_{\text{dense}}$ obtained in the previous section is a dense vector with spatiotemporal semantics, we can identify mobility patterns by clustering these dense vectors. In this section, the DBSCAN algorithm, a density-based clustering method, is applied to cluster dense trip vectors. For two trip vectors (i.e., $\mathbf{R}_{\text{dense}}$) containing mixed spatiotemporal information, the distance between them represents their spatiotemporal similarity. Additional details of the DBSCAN algorithm can be found in the study by [22]. One advantage of DBSCAN is that the number of clusters does not need to be manually specified in advance, which greatly reduces human intervention [40]. Instead, two parameters, the parameter of sample neighborhood size δ and the parameter of distance ε, are designed to describe the relationship between different samples to achieve clustering [31]. Here, we define a core sample to mean that there are at least δ other samples within the ε distance of a sample in the data set, and these samples are designated as neighbors of the core sample. For the trip vector, a core sample indicates that there are at least δ samples in the data set that have a spatiotemporal similarity less than ε. The flowchart of DBSCAN algorithm is shown in Figure 4. It can be seen that the key of the algorithm is to determine whether the sample is the core sample using the two parameters (δ and ε). Formally, assuming that the set of all dense trip vector is $\mathbf{R}_D$, given two dense trip vectors $\mathbf{R}^1_{\text{dense}}$ and $\mathbf{R}^2_{\text{dense}}$, $\mathbf{R}^1_{\text{dense}}, \mathbf{R}^2_{\text{dense}} \in \mathbf{R}_D$, the Manhattan distance is used to represent the difference in spatiotemporal semantics between them, which can be written as follows:

$$d_m\left(\mathbf{R}^1_{\text{dense}}, \mathbf{R}^2_{\text{dense}}\right) = \sum_{i=1}^{8} \left|\mathbf{R}^1_{\text{dense},i} - \mathbf{R}^2_{\text{dense},i}\right|, \tag{10}$$

where $\mathbf{R}^1_{\text{dense},i}$ and $\mathbf{R}^2_{\text{dense},i}$ represent the ith element in vector $\mathbf{R}^1_{\text{dense}}$ and $\mathbf{R}^2_{\text{dense}}$. The neighbor of the given trip vector $\mathbf{R}_{\text{dense}}$ can be expressed as follows:

$$\text{neighbor}\left(\mathbf{R}_{\text{dense}}\right) = \left\{\mathbf{R}^i_{\text{dense}} \middle| d_m\left(\mathbf{R}_{\text{dense}}, \mathbf{R}^x_{\text{dense}}\right) \leq \varepsilon, \mathbf{R}_{\text{dense}} \neq \mathbf{R}^x_{\text{dense}}, \forall \mathbf{R}^x_{\text{dense}} \in \mathbf{R}_D\right\}. \tag{11}$$

The condition that $\mathbf{R}_{\text{dense}}$ is the core sample can be expressed as follows:

$$\text{neighbor}\left(\mathbf{R}_{\text{dense}}\right) \geq \delta. \tag{12}$$

It needs to be clear that the values of parameters δ and ε need to be set in conjunction with the characteristics of the data set and the clustering target. Different values of the parameters have a significant impact on the clustering results. Here, two indicators are used to quantify algorithm performance, namely, the within-cluster sum of squared errors (SSE) and the silhouette coefficient (SC) [41]. Among them, SSE reflects the difference between different passengers who are identified as having the same mobility pattern. SSE in this paper can be calculated as follows:

$$\text{SSE} = \sum_{k=1}^{K} \sum_{m=1}^{M_k} \sum_{i=1}^{8} \left(\mathbf{R}^m_{\text{dense},i} - \mathbf{R}^{K\mu}_{\text{dense},i}\right)^2, \tag{13}$$

where K represents the number of clusters, M_k represents the number of samples in the kth cluster, $\mathbf{R}^m_{\text{dense},i}$ represents the ith element in the mth vector of the kth cluster, and $\mathbf{R}^{K\mu}_{\text{dense},i}$ represents the ith element in the center vector of the kth cluster. The smaller the value of SSE, the better the clustering performance. It means that passengers who are recognized as having the same mobility pattern have smaller identifiable differences, indicating that the pattern recognition is accurate. Besides, SC is a comprehensive index that combines cohesion and separation. Among them, the cohesion reflects the average difference between an individual passenger and other passengers identified as having the same mobility pattern. On the contrary, the separation means the smallest difference between the individual passenger and passengers with other mobility patterns. And then, SC in this paper can be expressed as follows:

$$\text{SC} = \frac{1}{\left|\mathbf{R}_D\right|} \cdot \sum_{m=1}^{\left|\mathbf{R}_D\right|} \frac{b_m - a_m}{\max\left(a_m, b_m\right)}, \tag{14}$$

where a_m reflects the degree of cohesion within a cluster, and b_m reflects the degree of separation between clusters. Specifically, for a trip vector $\mathbf{R}^m_{\text{dense}}$ ($\mathbf{R}^m_{\text{dense}} \in \mathbf{R}_D$), belonging to the kth cluster, the corresponding values of a_m and b_m can be calculated as follows:

$$a_m = \frac{1}{M_k - 1} \sum_{x=1, x \neq m}^{M_k} \sum_{i=1}^{8} \left(\mathbf{R}^m_{\text{dense},i} - \mathbf{R}^x_{\text{dense},i}\right)^2, \tag{15}$$

$$b_m = \min\left(b_{m,k'}\right), \quad k' \in (1, 2, \ldots, K), k' \neq k, \tag{16}$$

$$b_{m,k'} = \frac{1}{M_{k'} - 1} \sum_{x=1, x \neq m}^{M_{k'}} \sum_{i=1}^{8} \left(\mathbf{R}^m_{\text{dense},i} - \mathbf{R}^x_{\text{dense},i}\right)^2. \tag{17}$$

Indeed, from the above formulation, it is can be seen that $-1 \leq \text{SC} \leq 1$. If SC is close to 1, the data are well-clustered, indicating that the mobility pattern recognition is good. That is, the spatiotemporal characteristics of an individual passenger are highly similar to those of passengers in the same cluster. In contrast, passengers with different identified mobility patterns have significant differences in the spatiotemporal characteristics of travel. When SC is negative or even close to −1, it indicates that passengers with different travel spatiotemporal characteristics are identified as having the same pattern, which is obviously not ideal. In summary, the smaller SSE and larger SC (close to 1) characterize better mobility pattern recognition results.

4. Case Study and Applications

4.1. Case Description. A case study of Beijing metro network is presented to evaluate the proposed method. A total of 176.81 million passenger travel records from September to October 2018 are acquired to identify mobility patterns. Correspondingly, the POI data in Beijing during this period is also crawled from Amap.

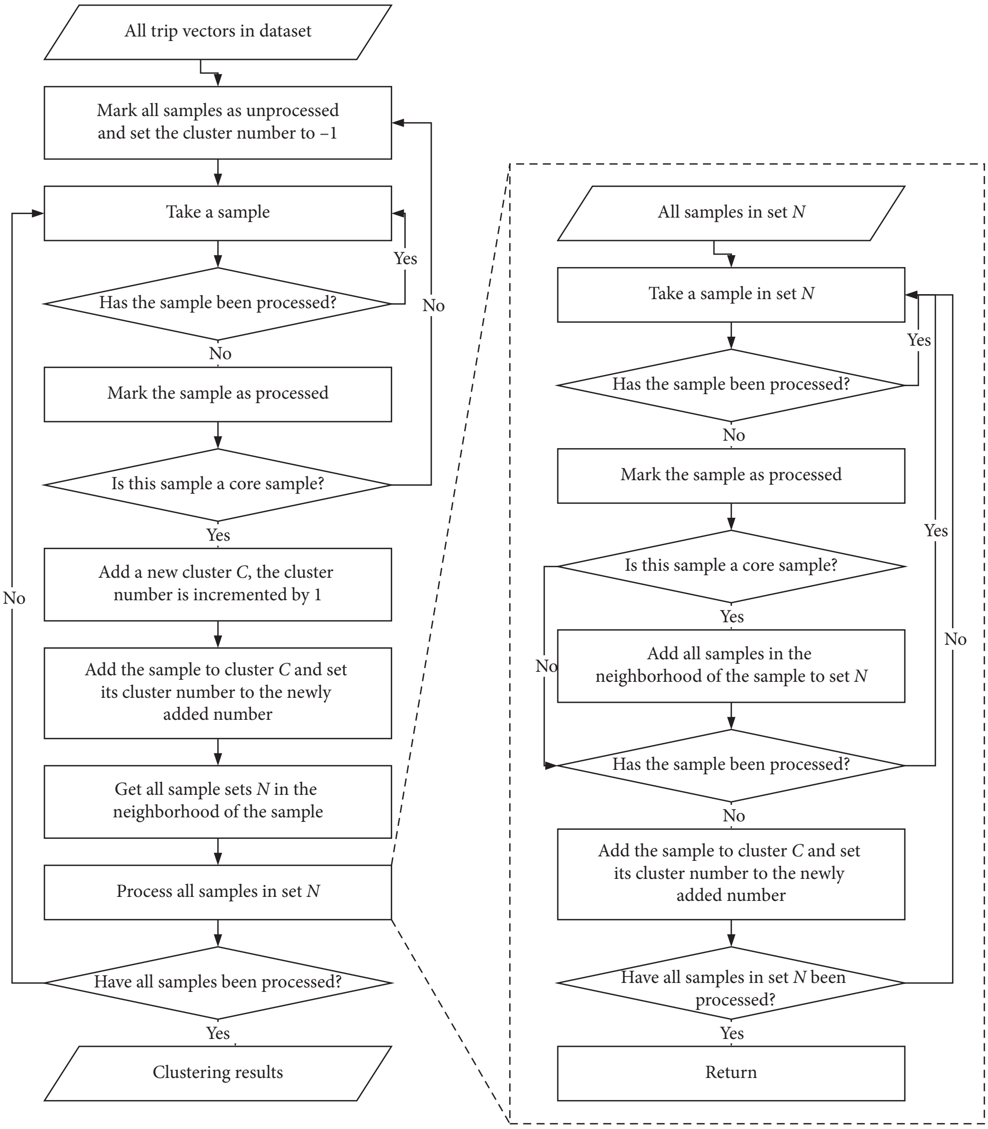

All trip vectors in dataset
Mark all samples as unprocessed and set the cluster number to −1
Take a sample
Has the sample been processed?
Yes
No
Mark the sample as processed
Is this sample a core sample?
No
Yes
Add a new cluster C, the cluster number is incremented by 1
No
Add the sample to cluster C and set its cluster number to the newly added number
Get all sample sets N in the neighborhood of the sample
Process all samples in set N
Have all samples been processed?
Yes
Clustering results
All samples in set N
Take a sample in set N
Yes
Has the sample been processed?
No
Mark the sample as processed
Yes
Is this sample a core sample?
Yes
No
No
Add all samples in the neighborhood of the sample to set N
Has the sample been processed?
No
Add the sample to cluster C and set its cluster number to the newly added number
Have all samples in set N been processed?
Yes
Return

Figure 4: DBSCAN algorithm flowchart.

First, the weighted POIs are generated by fusing multisource data from Lianjia and Arctiler using the method designed in Section 3.1. A total of 11,382 residential POIs were captured from Amap within the influence area of the metro station. Among them, 10927 residential POIs were successfully matched through the residential area data from Lianjia, indicating that the matching rate reached 96%. As for building data, a total of 6,887 buildings were captured from Amap. Among them, 6336 buildings were correctly matched through the Arctiler datasets, indicating a 92% match rate. It can be found that although there are some matching failures, the matching rates were higher than 90%, which proves that the proposed method can effectively weight the original POI data into weighted POIs. Residential POIs are used as examples to illustrate the advantages of weighted POI data, as shown in Figure 5. Among them, Figure 5(a) shows the distribution of Beijing metro stations, while Figures 5(b) and 5(c), respectively, show the distribution heat map of raw POIs and weighted POIs within 500 meters of metro stations. It can be seen that the residential POIs in Figure 5(b) are more evenly distributed and have a higher density in the urban center. On the contrary, the residential POIs in Figure 5(c) are concentrated in suburban areas in an extremely uneven manner. The reason for the above difference is that the residential POIs in the central area of the city are mainly hotels, villas, and residential buildings with few floors, while that in the suburban areas are mainly high-density, multistory residential communities. Furthermore, 4 high-density residential areas can be clearly observed in Figure 5(c), which are located in the north, east, and southwest of the city. Comparing existing studies, it can be found that the above regions correspond to Changping, Tongzhou, Fangshan, and Daxing, respectively [42–44]. The above areas have similar characteristics, such as low housing prices, high housing density, and a large number of commuters living in the area. It shows that weighted POI data can more accurately reflect the categories of land use around metro stations.

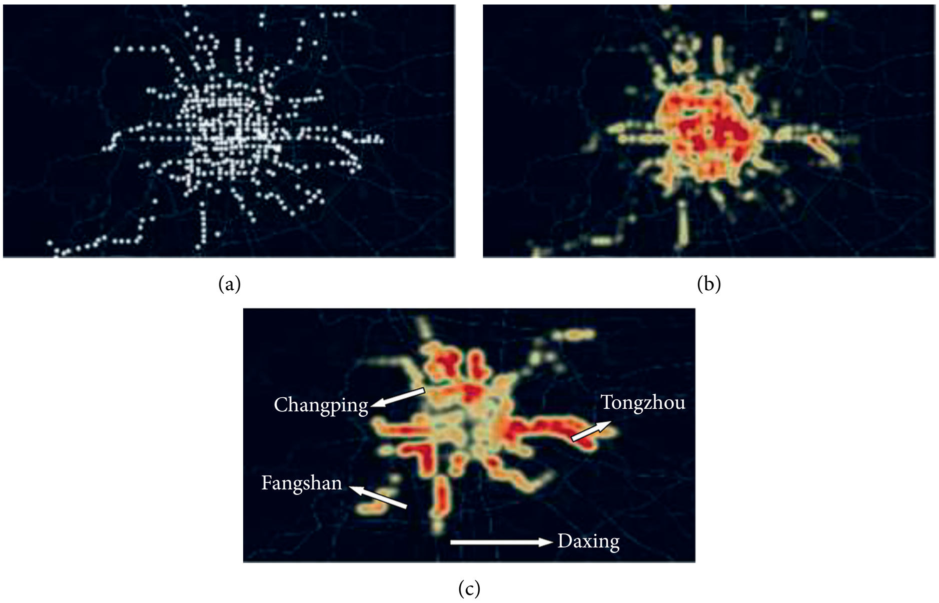


Figure 5: Comparison of weighted POIs and raw POIs. (a) The distribution of metro stations. (b) The distribution of raw POIs within 500 meters of metro stations. (c) The distribution of weighted POIs within 500 meters of metro stations.

Second, the spatiotemporal semantics are embedded using the SAE-based framework in Section 3.2. Figure 6 shows how the MSE changes with the number of iterations when training the SAE model. It can be seen that when the number of iterations reaches 40, the value of MSE remains stable. That is, SAE can encode the spatiotemporal features of the input trip records in a stable way, transforming the high-dimensional sparse vectors into low-dimensional dense vectors.

Third, the dense trip vectors are clustered using DBSCAN algorithm to realize the mobility pattern recognition. Since the number of clusters is not manually specified but is automatically generated according to the parameters δ and ε, it is necessary to check the number of clusters and algorithm performance corresponding to different values of parameters. This paper aims to identify passenger mobility patterns, so the number of clusters is required not to be too large (difficult to explain the potential activities of passengers) or too small (difficult to distinguish passenger categories) in order to balance practicality and interpretability. Through pre-experiments, we found that the number of clusters decreases as δ and ε increase. Further, when $\delta < 8$ and $\varepsilon < 7$, the number of clusters is verified to be greater than 30, which makes it difficult to accurately describe the potential activities represented by each mobility pattern. When $\delta > 18$ and $\varepsilon > 10$, the number of clusters is less than 3, which is obviously not conducive for our exploration of passenger mobility patterns. Therefore, the parameter value range is determined as: $\delta \in [8, 18]$ and $\varepsilon \in [7, 10]$. Figure 2 lists several results of the number of clusters and algorithm performance quantified by SSE and SC under different parameter values. It can be found that the value of SSE decreases with the increase of δ, and the influence of ε on SSE is limited. The relationship between SC and parameters is more complicated. Furthermore, the relationship between δ, ε, and SC is shown in Figure 7. From a global perspective, SC increases with the increase of parameters δ and ε. When δ reaches 16 and ε reaches 9.5, the value of SC decreases with the increase of parameters. Combining the above two indicators, a parameter combination of $\delta = 16$ and $\varepsilon = 9.5$ is selected. Herein, SSE = 23715 and SC = 0.815, showing good clustering performance.

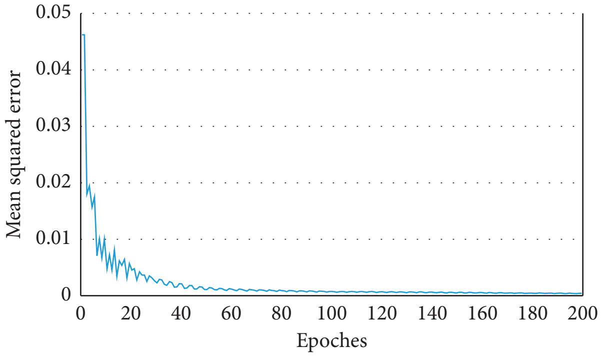


Figure 6: Changes in MSE during training SAE.

4.2. Results Analysis. The mobility pattern is recognized using the proposed method with the above parameters. Figure 8 shows the results when $\delta = 16$ and $\varepsilon = 9.5$. Each color represents a recognized mobility pattern and C1–C6 means the mobility features of cluster 1 to cluster 6. Among them, Figures 8(a) and 8(b) show the distribution of POI categories around the origin station and that around the destination station, which reveals the spatial features. The distributions of the start time of the day, the distribution of the day of week, and the distribution of travel time are presented in Figures 8(c)–8(e), respectively.

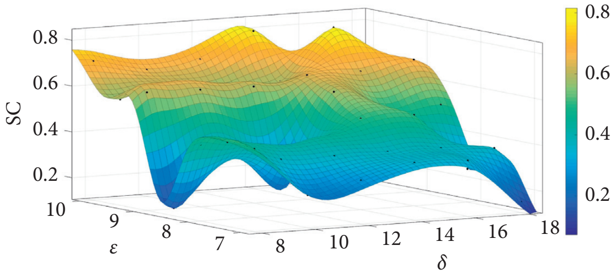


Figure 7: The value of SC corresponding to different parameters.

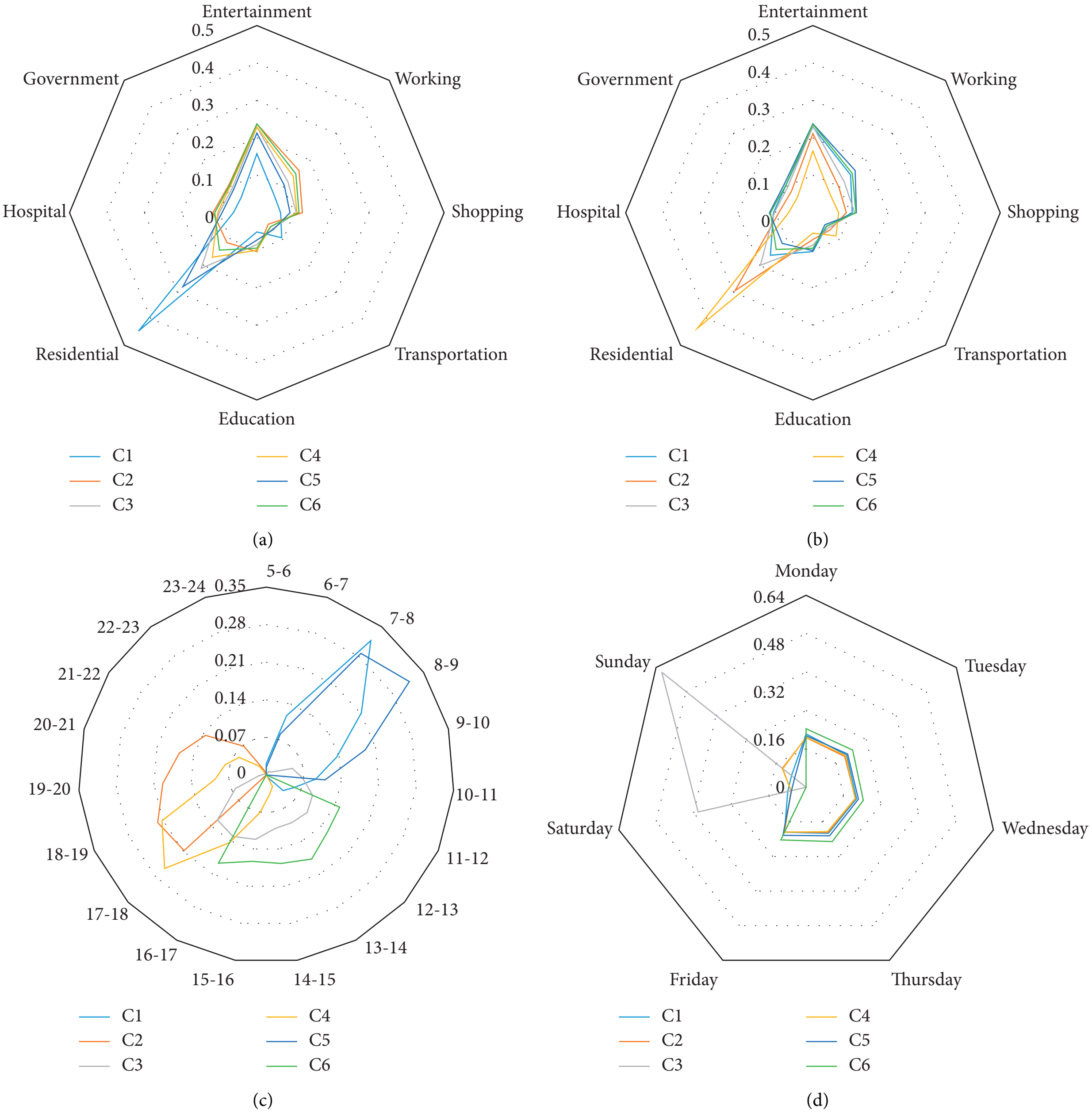
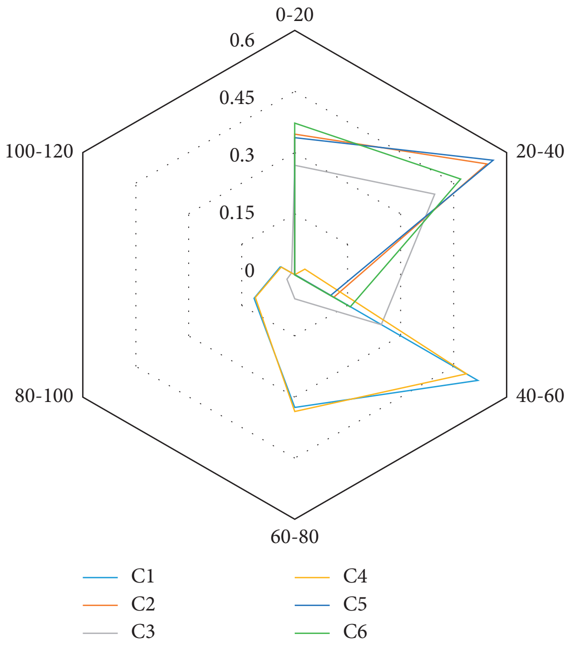


Figure 8: Continued.



(e)

Figure 8: Six mobility patterns recognized by the proposed method when $\delta = 16$ and $\varepsilon = 9.5$. (a) POI categories around the origin station. (b) POI categories around the destination station. (c) The start time of the day. (d) The day of week. (e) Travel time.

The characteristics of the six mobility patterns identified above are summarized, as shown in Table 2. Among them, C1 and C5 account for 35.808% (13.716% + 22.092%), representing the work-related mobility during the workdays. More specifically, C1 reveals long-distance working mobility, where the start time is between 7 and 8 am, and travel time is mainly 40–80 min. In contrast, C5 represents short-distance working, where the start time is between 7 and 9 am (later than the start time in C1), because travelers need to spend short travel time (mainly within 40 min). It can be found that although the temporal information is in line with the typical mobility patterns of commuters, the POIs around the destination station include multiple categories (not only working), such as entertainment, working, hospital, and shopping, which characterize the various possible work places of passengers. Besides, C3 (accounting for 13.908%) shows entertainment and shopping activities that mainly take place on weekends due to the large number of entertainment and shopping POIs surrounding the destination station. The start time of this type of mobility is between 9 am and 7 pm, and the travel time is within 60 min. Correspondingly, C2 and C4 account for 34.323% (19.817% + 14.506%), revealing the home-related mobility, most of which occurs on weekdays and Sundays. It can be seen that the destination POIs are mainly residential. The difference is that the start time of the mobility represented by C2 is all after 5 pm, while that represented by C4 is mainly concentrated between 5 pm and 7 pm. In C2 and C4, the various types of POIs (entertainment, working, shopping, etc.) around the origin station represent passengers at different working locations. Finally, C6 (accounting for 15.961%) represents a kind of mobility pattern. wherein it is difficult to directly identify the purpose of travel, where the origin location is mainly entertainment, shopping, and hospital POIs, the start time is between 11 am and 5 pm, and the travel time is within 40 min. The travel purpose of this pattern is difficult to be accurately identified, but it can be regarded as a short-distance travel that occurs during off-peak hours on weekdays.

The above analysis shows that mobility patterns related to working and home are the easiest to identify and explain, which is consistent with the conclusions of existing studies [23, 45, 46]. On the one hand, according to multidimensional temporal features, work-related mobility patterns can be divided into long-distance mobility and short-distance mobility. In this case, the number of short-distance travelers is 1.611 times (22.092%/13.716%) that of long-distance travelers, which shows that a large percentage of commuters work close to their places of residence. Nevertheless, there are still many commuters living far away from their working places, reflecting a serious imbalance between working and housing [43, 44]. On the other hand, home-related mobility patterns encompass more categories, because travelers can choose the time to go home more freely than the time to work. Taking C2 and C4 as examples, trips related to going home are clearly divided into two patterns. The start time of C2 is after 5 pm, and that of C4 is mainly between 5 and 7 pm. It can be inferred that the start time of the traveler's home trip is related to their work. In addition to working and going home, activities related to entertainment and shopping are displayed in C3. Most of them appear on weekends and their start time is between 9 am and 7 pm, which shows that passengers are more casual in choosing

Table 2: Characteristics of mobility patterns.

ID	Proportion (%)	Spatial features		Temporal features			Possible activity
		The origin station	The destination station	The start time	The day of week	Travel time	
C1	13.716	Mainly residential POIs	Mainly entertainment, working, hospital, and shopping POIs	Mainly 7–8	Mainly weekdays	Mainly 40 min–80 min	Working (long distance)
C2	19.817	Mainly entertainment, working, shopping, and education POIs	Mainly residential POIs	After 17	Weekdays and Sundays	Within 40 min	Home (short distance)
C3	13.908	Mainly residential, entertainment POIs	Mainly entertainment and shopping POIs	9–19	Mainly weekends	Within 60 min	Entertainment and shopping
C4	14.506	Mainly entertainment, working, and shopping POIs	Mainly residential POIs	Mainly 17–19	Weekdays and Sundays	Mainly 40 min– 80 min	Home (long distance)
C5	22.092	Mainly residential POIs	Mainly entertainment, working, shopping, and POIs	Mainly 7–9	Mainly weekdays	Mainly within 40 min	Working (short distance)
C6	15.961	Mainly entertainment, shopping, and hospital POIs	Mainly entertainment, shopping, and residential POIs	Mainly 11–17	Weekdays	Mainly within 40 min	Others

start time when engaging in entertainment and shopping activities. The above phenomenon is consistent with our empirical observation [18, 23]. It should be noted that the current analysis is based on the parameter settings of $\delta = 16$ and $\varepsilon = 9.5$. When the number of clusters decreases, several work-related patterns may be merged. Conversely, when the number of clusters increases, more mobility patterns may be found, but the difficulty of interpreting the pattern recognition results also increases.

It should be noted that sometimes the spatial information of the clustering results is confusing. For example, both clusters C1 and C3 have trips from residential POI to shopping POI. Nevertheless, C1 and C3 are interpreted as different potential activities (long-distance working/entertainment and shopping). This reflects the uncertainty of identifying passenger mobility patterns only through spatial information and the necessity of using spatiotemporal information jointly. For trips with the similar spatial information, temporal information can assist in inferring mobility patterns. Passengers who intend to shop are unlikely to choose to travel during the morning peak hours. They tend to choose off-peak hours to avoid crowded conditions and obtain higher travel comfort. It can be inferred that passengers in C1 who travel during the morning rush hours with shopping POIs as destinations are composed of most of the staff working in the mall and a small number of shoppers. Conversely, in C3, the potential activity of passengers traveling on weekends with shopping POIs as destinations is more likely to be shopping. When classifying a passenger's mobility pattern, the proposed embedding method can be used to embed the passenger's spatiotemporal information into a low-dimensional vector space. The distance between the embedded vector and the vector of each cluster center can be calculated to obtain the most likely mobility patterns and reduce the confusion caused by spatial information.

4.3. Sensitivity Analysis of Parameters. In this section, the sensitivity of parameters on the recognition results is analyzed. As shown in Table 3, the parameters of the clustering algorithm have a significant impact on the recognition performance. Here, we show the results when $\delta = 14$ and $\varepsilon = 10$ in Figure 9 and that when $\delta = 8$ and $\varepsilon = 10$ in Figure 10. In Figure 9, the trip vectors are divided into 3 patterns, SSE = 23647, and SC = 0.793. Obviously, it reveals the three most basic patterns of urban mobility: working, home, and others. Among them, C2 describes working-related trips, where the start time is mainly from 7 am to 9 am on weekdays, and the POIs around the origin station are mainly residential. Correspondingly, C3 represents trips related to going home, where the start time is mainly after 5 pm on weekdays, and the POIs around the destination station are dominated residential POIs. In addition, C3 represents trips that include entertainment, shopping, etc., where the start time is mainly distributed between 10 am and 5 pm on weekends. In Figure 10, the passenger trip vectors are identified as 11 clusters, SSE = 23133 and SC = 0.712. Compared with Figure 8, it can be seen that more passenger activities are identified.

Among them, C1, C9, and C11 are the three most easily explained patterns. They characterize working-related trips. In more detail, travel time of C1 is mainly 60–80 min, while that of C9 is within 40 min, and that of C11 is mainly 20–60 min. The travel time reflects the length of the journey. The three clusters C3, C8, and C10 represent home-related mobility. Their proportions are 13.606%, 7.478%, and 7.555%, respectively. In detail, the start time of C3 is mainly 5 pm–7 pm, while that of C8 is 7 pm–10 pm, and that of C10 is mainly 6 pm–8 pm. This shows that passengers are more flexible in time selection when going home. The remaining clusters represent mobility other than working and home, which are a refinement set of C3 and C6 in Table 2. Obviously, the mobility represented by these clusters is more dispersed in POI categories and more free

Table 3: The number of clusters and algorithm performance based on different parameters.

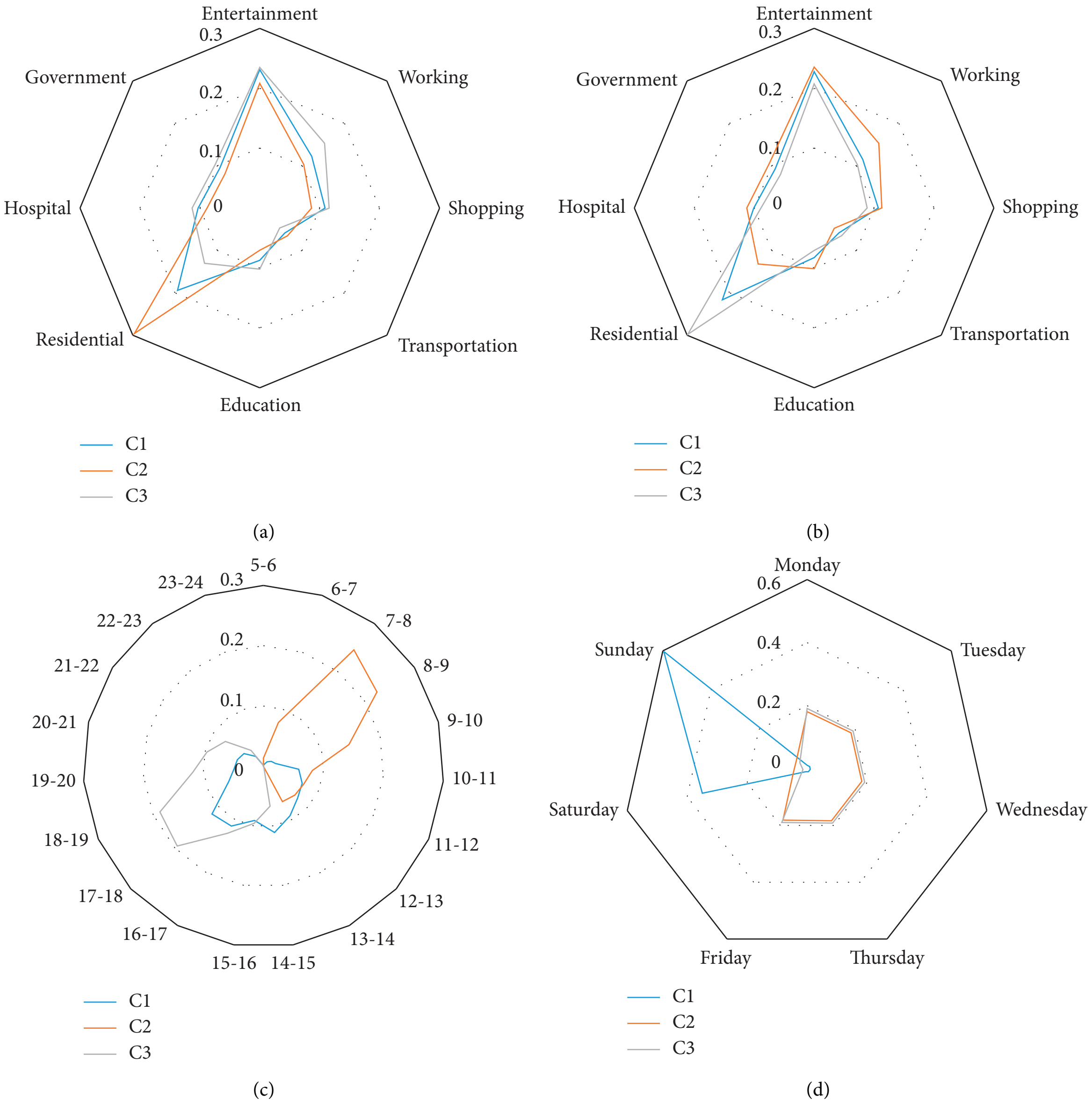
ID	δ	ε	K	SSE	SC
1	8	7	27	33136	0.466
2	8	7.5	29	33407	0.473
3	8	8	24	30895	0.439
4	8	8.5	18	34989	0.139
5	8	9	16	31018	0.621
6	8	9.5	15	31731	0.568
7	8	10	11	23133	0.712
8	10	7	24	28931	0.248
9	10	7.5	22	29168	0.38
10	10	8	19	27447	0.421
11	10	8.5	14	27925	0.361
12	10	9	13	27374	0.616
13	10	9.5	12	28828	0.621
14	10	10	9	27158	0.657
15	12	7	19	26676	0.295
16	12	7.5	17	26224	0.379
17	12	8	16	26786	0.424
18	12	8.5	11	26077	0.188
19	12	9	11	26775	0.614
20	12	9.5	10	26596	0.56
21	12	10	6	26638	0.686
22	14	7	18	25702	0.356
23	14	7.5	16	24084	0.381
24	14	8	12	25114	0.503
25	14	8.5	10	23429	0.597
26	14	9	9	23796	0.646
27	14	9.5	9	23889	0.596
28	14	10	3	23647	0.793
29	16	7	18	24333	0.348
30	16	7.5	14	22281	0.38
31	16	8	10	23117	0.502
32	16	8.5	10	23255	0.487
33	16	9	9	23760	0.646
34	**16**	**9.5**	**6**	**23715**	**0.815**
35	16	10	5	23568	0.596
36	18	7	17	22439	0.151
37	18	7.5	13	21892	0.364
38	18	8	10	23129	0.25
39	18	8.5	7	22538	0.507
40	18	9	7	23407	0.681
41	18	9.5	7	23436	0.629
42	18	10	6	23675	0.654

in the start time, which is in line with the diversified characteristics of weekend entertainment activities. Inevitably, as the number of clusters increases, the interpretability of the results is weakened. For example, there is no significant difference in the proportion of each category of POI around the origin station and the destination station in C4, which makes it difficult to find a known activity to explain its spatiotemporal characteristics. A feasible method is to investigate the purpose of the passengers in C4 to explain the above phenomenon. In summary, there must be a trade-off between the number of clusters and interpretability of the results.

4.4. Comparison of Methods. First, we compare the results with different vector forms. Here, sparse vectors (50 dimensions) and dense vectors (8 dimensions) are used to identify passenger mobility patterns, respectively. We examine the performance with different vector forms when the number of clusters is 6. It should be noted that after many pre-experiments, when sparse vectors are used and the number of clusters is 6, the input parameters are $\delta = 22$ and $\varepsilon = 72$. The comparison results are shown in Table 4. It can be seen that the calculation time using sparse vectors is much longer than that using dense vectors. This is because dense vectors need to consume less computing resources in the calculation process. In addition, compared to sparse vectors, using dense vectors can give better results, showing a smaller SSE and a larger SC. The reason is that the SAE-based embedding method efficiently extracts the spatiotemporal information in passenger travel records, which proves the necessity and superiority of embedding spatiotemporal semantics.

Next, we compare the performance of different methods. Here, two baseline methods are selected from the existing studies. The first one is a cluster-based method from literature [22]. Different from this paper, this method aims to mine the spatiotemporal travel patterns from the long-term historical travel database, whereas OD stations are regarded as spatial features and the timestamps of entering and exiting stations are regarded as temporal features. The second baseline method is a topic model based on LDA from literature [23]. In this model, the four features are considered to describe a passenger trip—they are the location (station), start time of day, start day of week, and the duration. It should be noted that this model is a "soft-cluster" method, where a probability distribution is used to quantify the relationship between a trip and mobility patterns.

Due to the lack of real activity labels for passenger travel records, it is challenging to quantify and compare the performance of various methods in mobility pattern recognition in terms of accuracy. One way to deal with this problem is to design a stated preference (SP) survey to determine the actual travel purpose of passengers, which can be utilized as a benchmark to calculate the accuracy of the mobility pattern recognition results [47]. Nevertheless, SP surveys often require huge manpower and material resources, especially in large-scale analysis. In this section, a compromise method is adopted to evaluate the performance of models by using the SSE calculated by equation (13) and the SC calculated by equation (14). These two indicators can measure the ability of the pattern recognition results to characterize the distribution of the data, evaluating the models without real activity labels [23]. Based on the data in Section 4.1, the number of clusters is set to 3, 6, and 11 respectively, and the above two methods are used to recognize mobility patterns. Figure 11 shows the values of the two indicators (SSE and SC) corresponding to the results obtained by different methods, in which the K represents the number of clusters. It can be found that when the number of cluster is 3 and 6, the SSE of baseline 2 and that of the proposed method are relatively small, while that of baseline 1 is larger. When the number of clusters is 11, the three methods have comparable SSE. This means a significant intra-cluster difference of identified mobility patterns when the OD stations are regarded as the spatial



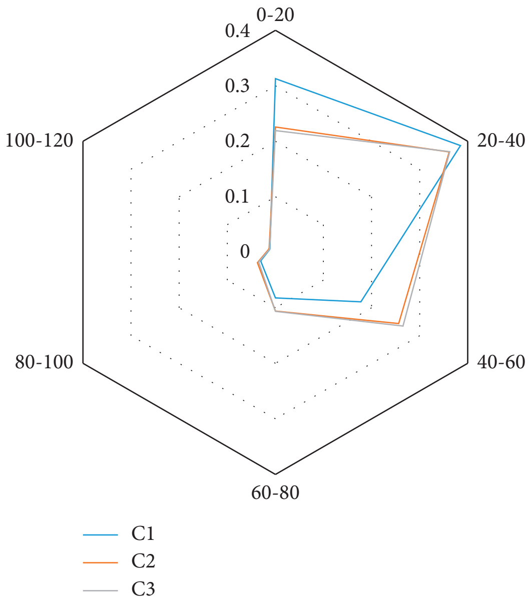
(a)

(b)

(c)

(d)

FIGURE 9: Continued.



(e)

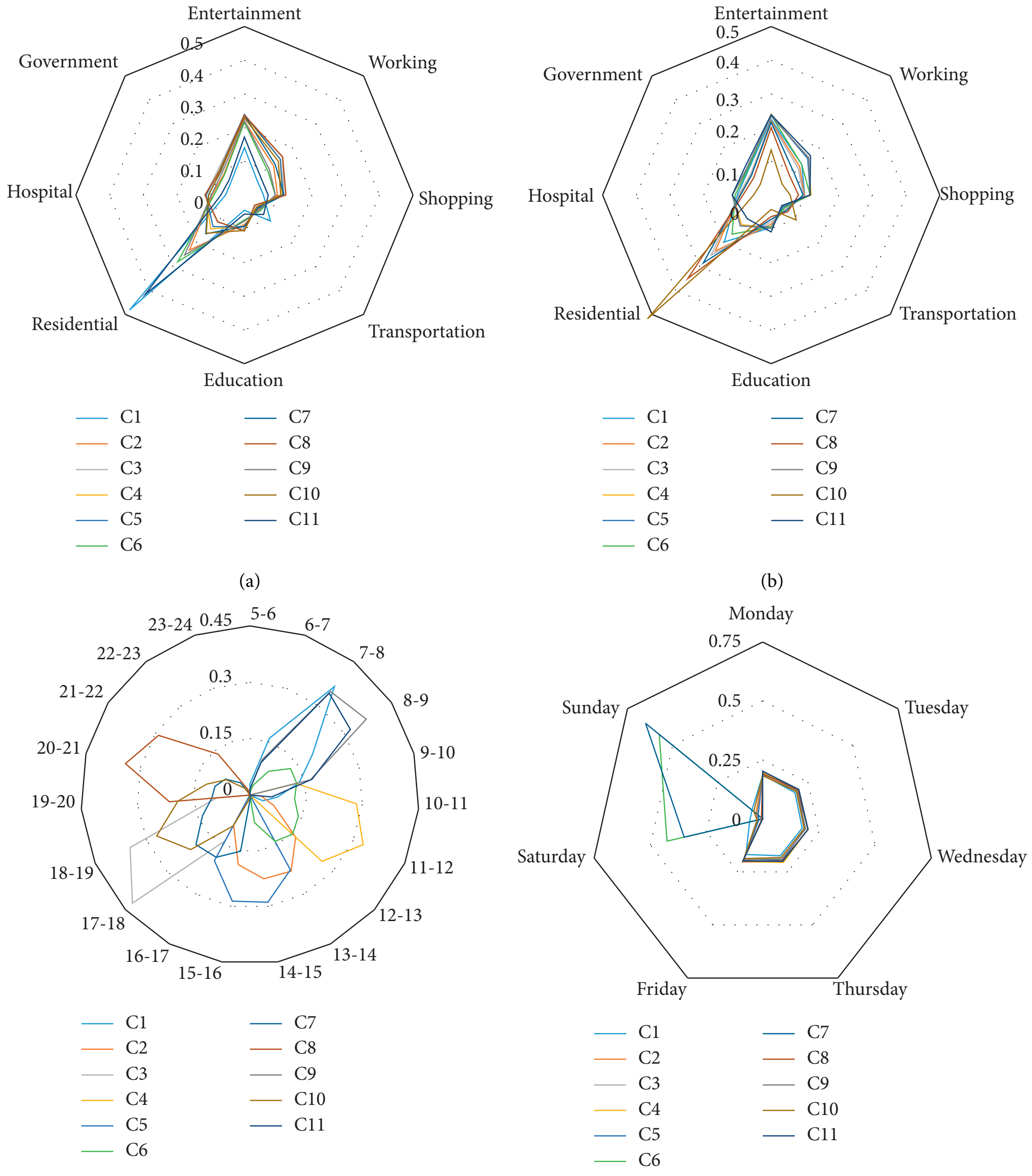
Figure 9: Three mobility patterns recognized by the proposed method when $\delta = 14$ and $\varepsilon = 10$. (a) POI categories around the origin station. (b) POI categories around the destination station. (c) The start time of the day. (d) The day of week. (e) Travel time.

features. Conversely, the proposed method and baseline 2 perform better in terms of SSE. Nevertheless, with the same number of clusters, the proposed method has a larger SC value than baseline 2. This shows that baseline 2 fails to distinguish the trips in different patterns well. In summary, the proposed method performs well in mobility analysis, which illustrates the necessity of using weighted POIs based on multisource data to characterize spatial attributes and the superiority of using coding-based methods to vectorize passenger trips.

4.5. Applications Based on Passenger Mobility Patterns. The ultimate goal of mobility pattern recognition is to accurately grasp the characteristics of passenger needs and assist subway operators and managers to provide passengers with high-quality travel services. As described in Section 4.2, with the help of mobility pattern recognition, the time preferences, start location preferences, and the attributes of potential activities of different types of passengers can be explored. Furthermore, when a certain passenger's historical travel data are given, his/her mobility mode type can be calculated through similarity calculation, individual travel preferences can be estimated, and demand characteristics can be clarified. Based on this, it has become possible to provide personalized services according to individual travel needs.

On the one hand, individual mobility pattern helps generate more accurate personalized passenger guidance strategies. In traditional practice, metro operators empirically recommend the route with the shortest travel time or the lowest travel cost to passengers. Nevertheless, existing studies have shown that passengers with different travel purposes pay different attention to different factors [47, 48]. For example, commuters may be more concerned about travel time reliability. On the contrary, travelers do not have high requirements for travel time reliability but are more concerned about the comfort of travel. The identification and analysis of mobility patterns can help provide personalized guidance strategies.

On the other hand, mobility pattern recognition can be used as a powerful tool to guide business applications. Here, the applications in advertising and Mobility-as-a-Service (MaaS) design and promotion are introduced. For advertisers, it would be wise to consider the passenger demand of the station when placing advertisements at a designated station, which can be obtained through the research of this paper. Related researches have been conducted in recent years to support mobility-pattern-based advertising [21, 49]. For example, it is obvious that in stations where many commuters live in the surrounding area, recruitment and job-hunting advertisements are very competitive. Besides, as a technological innovation with the potential to revolutionise the urban mobility paradigm, MaaS is emerging and closely related to mobility pattern recognition. MaaS is a service offered to the user in a single mobile app platform, which integrates all aspects of the travel experience, including booking, payment, and information, both before and during the trip [50]. The latest research shows that understanding passengers' mobility patterns and expectations is key for designing successful MaaS technologies [51]. And then, researches also show that willingness to use MaaS is strongly correlated with age and lifecycle stage, which can be identified by the proposed method in this paper [52]. For example, young individuals who are employed full-time are most likely to use MaaS.



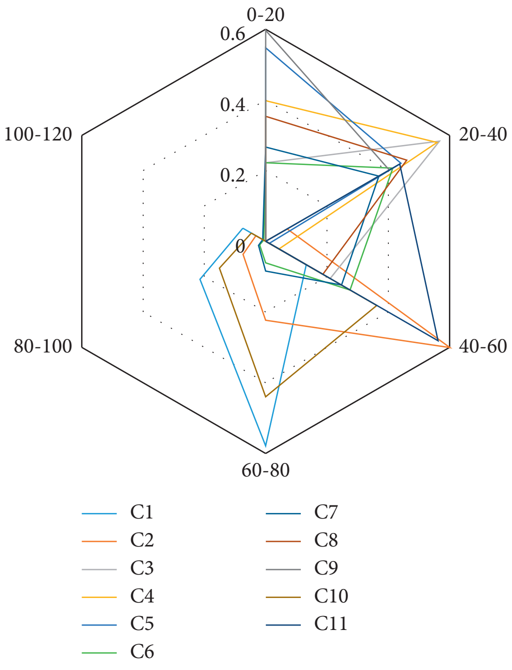
(a)

(b)

(c)

(d)

Figure 10: Continued.



(e)

FIGURE 10: Eleven mobility patterns recognized by the proposed method when $\delta = 8$ and $\varepsilon = 10$. (a) POI categories around the origin station. (b) POI categories around the destination station. (c) The start time of the day. (d) The day of week. (e) Travel time.

TABLE 4: Comparison of results with different vector forms.

Vector form	Computing time (second)	SSE	SC
Sparse vector	2674	28706	0.602
Dense vector	144	23715	0.815

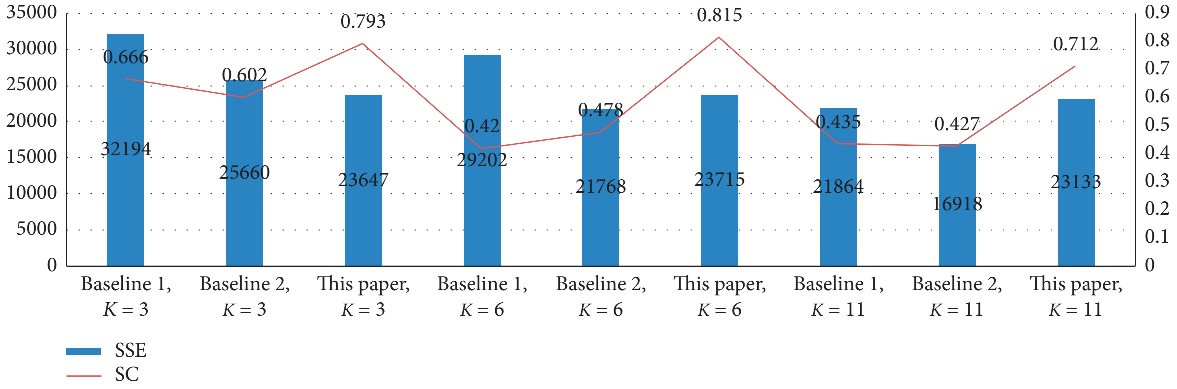


FIGURE 11: Comparison of methods.

It should be noted that mobility pattern recognition also has important applications in the prevention and control of epidemic spreading and the assessment of social and economic development. For details, please refer to references [10, 27].

5. Conclusions and Discussion

This paper presents a SAE-based unsupervised learning framework to explore the potential of AFC data in recognizing passenger mobility patterns. The proposed model converts the travel records of passengers into trip vectors in an embedded manner to facilitate large-scale pattern recognition. Each trip vector contains spatial attributes (POIs around the origin station, POIs around the destination station) and temporal attributes (start time, day of the week, and travel time), which enhance the interpretability of the mobility analysis results. Specifically, the spatial characteristics are obtained through the fusion of multisource, geo-based data. A density-based clustering algorithm is introduced to group the trip vectors into multiple clusters to realize the mobility pattern recognition. A case of Beijing metro network is used to verify the feasibility of the above methods. In this case, six typical mobility patterns are identified, two of which are related to working (accounting for 36.702%), three of which are related to home (accounting for 46.057%), and one of which is related to entertainment and studying (accounting for 17.242%), revealing the mobility distribution characteristics of Beijing metro passengers. Furthermore, the sensitivity analysis of the parameters is done. It is found that as the number of clusters in the results increases, the identified mobility patterns can reflect more detailed passenger activity characteristics and at the same time have greater inexplicability. The comparison with the other two baseline methods proves that the proposed method can better explore the passenger mobility patterns based on multisource data than the existing methods. This research provides a way of embedding complex, multisource, and different-dimensional spatiotemporal information into dense trip vectors, which is suitable for large-scale calculations to identify mobility patterns.

Admittedly, the proposed method still has several limitations that can be considered in the future works. First, geographic information needs to be processed more finely. This paper divides the captured POIs into 8 categories as shown in Table 1, and each category contains multiple subcategories. There may be great differences between subcategories. For example, Card & Chess Room and Camping Site are considered the same category (entertainment) in this paper. In fact, these two kinds of POIs can be treated separately as indoor entertainment and outdoor sports, which helps to discover more detailed features of passenger activities. Second, the dependence between multiple trips of a passenger needs to be considered. This paper only embeds the spatiotemporal features of the current trip into the dense vector, and does not consider the previous and subsequent trips, which limits the application of the proposed method in the generation of passenger activity chains and the prediction of trips [32]. For example, a point of view is widely agreed that, due to geographical constraints, the origin station of the current trip is likely to be the destination station of the previous trip. It means that considering the information of previous and subsequent trips to complete the embedding of the current trip has potential application value. Third, although the validity of the matching between the selected multiple data sources is acceptable, there are still some matching failures. The selection of data sources with better matching is worth exploring to improve the proposed data fusion method. This will be the focus of the future studies.

Data Availability

All data, models, or codes that support the findings of this study are available from the corresponding author upon reasonable request.

Conflicts of Interest

The authors declare no conflicts of interest.

Authors' Contributions

Chao Yu was in charge of conceptualization, writing the original draft, providing software, and visualization. Haiying Li was concerned with methodology and project administration. Xinyue Xu was involved in conceptualization, methodology, reviewing, and editing. Jun Liu did supervision and funding acquisition. Jianrui Miao was responsible for methodology. Yitang Wang and Qi Sun contributed to data curation.

Acknowledgments

This work was financially supported by the National Natural Science Foundation of China (Grant No. 71871012) and the State Key Lab of Rail Traffic Control and Safety of China (Grant No. RCS2020ZT005).

References

[1] A. Alsger, A. Tavassoli, M. Mesbah, L. Ferreira, and M. Hickman, "Public transport trip purpose inference using smart card fare data," *Transportation Research Part C: Emerging Technologies*, vol. 87, pp. 123–137, 2018.

[2] S. Wang, L. Li, W. Ma, and X. Chen, "Trajectory analysis for on-demand services: a survey focusing on spatial-temporal demand and supply patterns," *Transportation Research Part C: Emerging Technologies*, vol. 108, pp. 74–99, 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, 2020.

[4] H. Li, Y. Wang, X. Xu, L. Qin, and H. Zhang, "Short-term passenger flow prediction under passenger flow control using a dynamic radial basis function network," *Applied Soft Computing*, vol. 83, Article ID 105620, 2019.

[5] X. Xu, H. Li, J. Liu, B. Ran, and L. Qin, "Passenger flow control with multi-station coordination in subway networks: algorithm development and real-world case study," *Transportmetrica B: Transport Dynamics*, pp. 1–27, 2018.

[6] K. Farrahi and D. Gatica-Perez, "Discovering routines from large-scale human locations using probabilistic topic models," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 1, pp. 1–27, 2011.

[7] H. Cao, F. Xu, J. Sankaranarayanan, Y. Li, and H. Samet, "Habit2vec: trajectory semantic embedding for living pattern recognition in population," *IEEE Transactions on Mobile Computing*, vol. 19, no. 5, pp. 1096–1108, 2020.

[8] M. Allahviranloo and W. Recker, "Daily activity pattern recognition by using support vector machines with multiple classes," *Transportation Research Part B: Methodological*, vol. 58, pp. 16–43, 2013.

[9] S. Hasan and S. V. Ukkusuri, "Urban activity pattern classification using topic models from online geo-location data," *Transportation Research Part C: Emerging Technologies*, vol. 44, pp. 363–381, 2014.

[10] J. Blumenstock, G. Cadamuro, and R. On, "Predicting poverty and wealth from mobile phone metadata," *Science*, vol. 350, no. 6264, pp. 1073–1076, 2015.

[11] L. Alexander, S. Jiang, M. Murga, and M. C. González, "Origin-destination trips by purpose and time of day inferred from mobile phone data," *Transportation*

Research Part C: Emerging Technologies, vol. 58, pp. 240–250, 2015.

[12] Y. Xu, A. Belyi, I. Bojic, and C. Ratti, "Human mobility and socioeconomic status: analysis of Singapore and Boston," *Computers, Environment and Urban Systems*, vol. 72, pp. 51–67, 2018.

[13] 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.

[14] L.-M. Kieu, A. Bhaskar, and E. Chung, "A modified density-Based Scanning Algorithm with Noise for spatial travel pattern analysis from Smart Card AFC data," *Transportation Research Part C: Emerging Technologies*, vol. 58, pp. 193–207, 2015.

[15] G. Han and K. Sohn, "Activity imputation for trip-chains elicited from smart-card data using a continuous hidden Markov model," *Transportation Research Part B: Methodological*, vol. 83, pp. 121–135, 2016.

[16] L. Sun and K. W. Axhausen, "Understanding urban mobility patterns with a probabilistic tensor factorization framework," *Transportation Research Part B: Methodological*, vol. 91, pp. 511–524, 2016.

[17] M. K. El Mahrsi, E. Come, L. Oukhellou, and M. Verleysen, "Clustering smart card data for urban mobility analysis," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 3, pp. 712–728, 2017.

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

[19] J. B. Gordon, H. N. Koutsopoulos, and N. H. M. Wilson, "Estimation of population origin-interchange-destination flows on multimodal transit networks," *Transportation Research Part C: Emerging Technologies*, vol. 90, pp. 350–365, 2018.

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

[21] D. Zhuang, S. Hao, D.-H. Lee, and J. G. Jin, "From compound word to metropolitan station: semantic similarity analysis using smart card data," *Transportation Research Part C: Emerging Technologies*, vol. 114, pp. 322–337, 2020.

[22] L. M. Kieu, A. Bhaskar, and E. Chung, "Passenger segmentation using smart card data," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 3, pp. 1537–1548, 2015.

[23] Z. Zhao, H. N. Koutsopoulos, and J. Zhao, "Discovering latent activity patterns from transit smart card data: a spatiotemporal topic model," *Transportation Research Part C: Emerging Technologies*, vol. 116, Article ID 102627, 2020.

[24] Z. Chen, Y. Zhang, C. Wu, and B. Ran, "Understanding individualization driving states via latent dirichlet allocation model," *IEEE Intelligent Transportation Systems Magazine*, vol. 11, no. 2, pp. 41–53, 2019.

[25] Z. Chen, H. Cao, H. Wang, F. Xu, V. Kostakos, and Y. Li, "Will you come back/check-in again? understanding characteristics leading to urban revisitation and Re-check-in," *Proceedings of the ACM Interactive, Mobile, Wearable Ubiquitous Technol.*vol. 4, no. 3, 2020.

[26] J. Wang, X. Kong, F. Xia, and L. Sun, "Urban human mobility," *ACM SIGKDD Explorations Newsletter*, vol. 21, no. 1, pp. 1–19, 2019.

[27] H. Barbosa, M. Barthelemy, G. Ghoshal et al., "Human mobility: models and applications," *Physics Reports*, vol. 734, pp. 1–74, 2018.

[28] Y. Zhang, N. Sari Aslam, J. Lai, and T. Cheng, "You are how you travel: a multi-task learning framework for geodemographic inference using transit smart card data," *Computers, Environment and Urban Systems*, vol. 83, Article ID 101517, 2020.

[29] M. Ren, Y. Lin, M. Jin, Z. Duan, Y. Gong, and Y. Liu, "Examining the effect of land-use function complementarity on intra-urban spatial interactions using metro smart card records," *Transportation*, vol. 47, no. 4, pp. 1607–1629, 2020.

[30] Y. Gong, Y. Lin, and Z. Duan, "Exploring the spatiotemporal structure of dynamic urban space using metro smart card records," *Computers, Environment and Urban Systems*, vol. 64, pp. 169–183, 2017.

[31] E. Schubert, J. Sander, M. Ester, H. P. Kriegel, and X. Xu, "DBSCAN revisited, revisited: why and how you should (still) use DBSCAN," *ACM Trans. Database Syst.*vol. 42, no. 3, 2017.

[32] Z. Zhao, H. N. Koutsopoulos, and J. Zhao, "Individual mobility prediction using transit smart card data," *Transportation Research Part C: Emerging Technologies*, vol. 89, pp. 19–34, 2018.

[33] C. Yu, H. Li, X. Xu, and J. Liu, "Data-driven approach for solving the route choice problem with traveling backward behavior in congested metro systems," *Transportation Research Part E: Logistics and Transportation Review*, vol. 142, Article ID 102037, 2020.

[34] F. Jia, H. Li, X. Jiang, and X. Xu, "Deep learning-based hybrid model for short-term subway passenger flow prediction using automatic fare collection data," *IET Intelligent Transport Systems*, vol. 13, no. 11, pp. 1708–1716, 2019.

[35] X. Yuan, B. Huang, Y. Wang, C. Yang, and W. Gui, "Deep learning-based feature representation and its application for soft sensor modeling with variable-wise weighted SAE," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 7, pp. 3235–3243, 2018.

[36] P. Baldi, "Autoencoders, unsupervised learning, and deep architectures," *ICML Unsupervised and Transfer Learning*, vol. 27, pp. 37–50, 2012.

[37] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, "Greedy layer-wise training of deep networks," *Advances in Neural Information Processing Systems*, vol. 1, pp. 153–160, 2007.

[38] M. C. González, C. A. Hidalgo, and A.-L. Barabási, "Understanding individual human mobility patterns," *Nature*, vol. 453, no. 7196, pp. 779–782, 2008.

[39] C. Song, T. Koren, P. Wang, and A.-L. Barabási, "Modelling the scaling properties of human mobility," *Nature Physics*, vol. 6, no. 10, pp. 818–823, 2010.

[40] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*, Menlo Park, CA, USA, August 1996.

[41] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987.

[42] M. Zhang, S. He, and P. Zhao, "Revisiting inequalities in the commuting burden: institutional constraints and job-housing relationships in Beijing," *Journal of Transport Geography*, vol. 71, pp. 58–71, 2018.

[43] P. Qin and L. Wang, "Job opportunities, institutions, and the jobs-housing spatial relationship: case study of Beijing," *Transport Policy*, vol. 81, pp. 331–339, 2019.

[44] P. Zhao, D. Liu, Z. Yu, and H. Hu, "Long commutes and transport inequity in China's growing megacity: new evidence from Beijing using mobile phone data," *Travel Behaviour and Society*, vol. 20, pp. 248–263, 2020.

[45] H. Ballis and L. Dimitriou, "Revealing personal activities schedules from synthesizing multi-period origin-destination matrices," *Transportation Research Part B: Methodological*, vol. 139, pp. 224–258, 2020.

[46] E. Chen, Z. Ye, C. Wang, and W. Zhang, "Discovering the spatio-temporal impacts of built environment on metro ridership using smart card data," *Cities*, vol. 95, Article ID 102359, 2019.

[47] N. C. Iraganaboina, T. Bhowmik, S. Yasmin, N. Eluru, and M. A. Abdel-Aty, "Evaluating the influence of information provision (when and how) on route choice preferences of road users in Greater Orlando: application of a regret minimization approach," *Transportation Research Part C: Emerging Technologies*, vol. 122, Article ID 102923, 2021.

[48] A. Ceder and Y. Jiang, "Route guidance ranking procedures with human perception consideration for personalized public transport service," *Transportation Research Part C: Emerging Technologies*, vol. 118, Article ID 102667, 2020.

[49] H. Yin, B. Cui, Z. Huang, W. Wang, X. Wu, and X. Zhou, "Joint modeling of users' interests and mobility patterns for point-of-interest recommendation," in *Proceedings of the 2015 ACM Multimedia Conference*, Portland, OR, USA, March 2015.

[50] M. J. Alonso-González, S. Hoogendoorn-Lanser, N. van Oort, O. Cats, and S. Hoogendoorn, "Drivers and barriers in adopting Mobility as a Service (MaaS) – a latent class cluster analysis of attitudes," *Transportation Research Part A*, vol. 132, pp. 378–401, 2020.

[51] I. Lopez-Carreiro, A. Monzon, E. Lopez, and M. E. Lopez-Lambas, "Urban mobility in the digital era: an exploration of travellers' expectations of MaaS mobile-technologies," *Technology in Society*, vol. 63, Article ID 101392, 2020.

[52] A. Vij, S. Ryan, S. Sampson, and S. Harris, "Consumer preferences for Mobility-as-a-Service (MaaS) in Australia," *Transportation Research Part C: Emerging Technologies*, vol. 117, Article ID 102699, 2020.

Hindawi
Journal of Advanced Transportation
Volume 2021, Article ID 9945398, 13 pages
https://doi.org/10.1155/2021/9945398

Research Article

A Two-Level Model for Traffic Signal Timing and Trajectories Planning of Multiple CAVs in a Random Environment

Yangsheng Jiang,[1,2,3] **Bin Zhao**,[1,2] **Meng Liu,**[1,2] **and Zhihong Yao**[1,2,3,4]

[1]*School of Transportation and Logistics, Southwest Jiaotong University, Chengdu 610031, Sichuan, China*
[2]*National Engineering Laboratory of Integrated Transportation Big Data Application Technology, Southwest Jiaotong University, Chengdu 610031, Sichuan, China*
[3]*National United Engineering Laboratory of Integrated and Intelligent Transportation, Southwest Jiaotong University, Chengdu 610031, Sichuan, China*
[4]*Institute of System Science and Engineering, Southwest Jiaotong University, Chengdu 610031, Sichuan, China*

Correspondence should be addressed to Zhihong Yao; zhyao@my.swjtu.edu.cn

Received 10 March 2021; Revised 29 March 2021; Accepted 19 April 2021; Published 27 April 2021

Academic Editor: Linchao Li

Connected and automated vehicles (CAVs) trajectories not only provide more real-time information by vehicles to infrastructure but also can be controlled and optimized, to further save travel time and gasoline consumption. This paper proposes a two-level model for traffic signal timing and trajectories planning of multiple connected automated vehicles considering the random arrival of vehicles. The proposed method contains two levels, i.e., CAVs' arrival time and traffic signals optimization, and multiple CAVs trajectories planning. The former optimizes CAVs' arrival time and traffic signals in a random environment, to minimize the average vehicle's delay. The latter designs multiple CAVs trajectories considering average gasoline consumption. The dynamic programming (DP) and the General Pseudospectral Optimal Control Software (GPOPS) are applied to solve the two-level optimization problem. Numerical simulation is conducted to compare the proposed method with a fixed-time traffic signal. Results show that the proposed method reduces both average vehicle's delay and gasoline consumption under different traffic demand significantly. The average reduction of vehicle's delay and gasoline consumption are 26.91% and 10.38%, respectively, for a two-phase signalized intersection. In addition, sensitivity analysis indicates that the minimum green time and free-flow speed have a noticeable effect on the average vehicle's delay and gasoline consumption.

1. Introduction

Traffic congestion has become a common traffic phenomenon in many cities [1]. In the United States, the transportation sector consumed about 143 billion gallons of gasoline in 2017 [2]. Moreover, traffic congestion leads to additional transportation emissions and travel delays. In 2017, due to traffic congestion, drivers in the United States waste an average of 41 hours per year during peak hours [3]. Therefore, it is urgent to save gasoline consumption and travel time in cities [4, 5].

As one of the effective methods to alleviate urban traffic congestion [6], traffic signal control [7] first appeared in London, England, in 1868. Currently, traffic signal control mainly consists of three strategies: fixed-time control, vehicle-actuated control, and traffic signal adaptive control. These strategies allocate space-time right of way to vehicles in different conflict directions to resolve traffic flow conflicts at intersections [8]. However, these control strategies rely on traffic data from infrastructure-based vehicle detection systems, such as loop detectors, radar, or cameras [9–11]. Infrastructure-based vehicle detection systems only provide limited discrete data, and their installation and maintenance costs are considerably high [9]. Recently, with the development of wireless communication and automatic driving technologies, CAVs can realize the information exchange between vehicles and infrastructure (i.e., traffic signal equipment) [12, 13]. Therefore, traffic signals and vehicle trajectories can be optimized and designed for connected automated vehicles (CAVs) to improve traffic efficiency and save gasoline consumption.

Many works have been conducted to search the optimal traffic signal and vehicle trajectories in the CAVs environment [14]. These works can be divided threefold. Firstly, a large number of signal control algorithms were proposed to optimize traffic signals with CAVs data [12, 15–20]. Secondly, many studies designed vehicle trajectories for CAVs to save gasoline consumption [21–26]. Thirdly, several methods focused on optimizing traffic signals and CAVs' trajectories to save travel time and gasoline consumption [8, 27–29].

However, there are several limitations to current integrated optimization methods. First, Feng et al. [8] and Yu et al. [29] only optimized the leading vehicle trajectory of a platoon and a car-following model that calculates the other vehicles' trajectories. Second, Xu et al. [28] proposed a vehicle trajectory designing model that considered a safe front vehicle distance. Still, they did not consider optimizing the trajectory of all CAVs at the same time. Therefore, this study would fill in this gap by showing a two-level model for traffic signal timing and trajectories planning of multiple connected automated vehicles considering the random arrival of vehicles.

The contribution of this paper consists of extending the optimal framework in Feng et al. [8]. First, instead of optimizing traffic signals by dynamic programming [8], we formulate an optimal arrival time calculation model for each CAV based on traffic signal timing and optimize traffic signals and vehicles' arrival time for random arrival CAVs to minimize average vehicle's delay. Second, unlike Feng et al. [8] and Yu et al. [29], only optimizing the leading vehicle trajectory of a platoon, the other vehicle trajectories are generated by a car-following model. Here, we proposed a multiple CAVs trajectories planning model, which is solved by the GPOPS [30]. Compared with Feng et al. [8], Yu et al. [29], and Xu et al. [28], the proposed model can optimize the trajectories of multiple CAVs at the same time. Third, we develop a two-level optimization framework and algorithm. Finally, we design the numerical examples and investigate the influence of critical parameters on the proposed method's performance.

The remainder of the paper is organized as follows. Section 2 reviews the research on traffic signal and trajectory optimization. Section 3 introduces some assumptions, two-level model, and solution algorithm. Section 4 presents numerical experiments, discussions, and sensitivity analysis. Finally, conclusions and recommendations are delivered in Section 5.

2. Literature Review

Connected and automated vehicles (CAVs) have great potential in improving traffic efficiency and reducing traffic congestion and have gained a wide application in the transportation field during the last decade [31]. These applications mainly focus on CAV-based trajectories planning [22, 23, 25, 26, 32–34] and CAV-based signal timing optimization [9, 12, 16, 35] and even further to design traffic signals and CAVs trajectories simultaneously [8, 27, 29, 34, 36, 37]. These studies showed that CAVs applications in trajectories planning and signal timing optimization could further reduce gasoline consumption, pollutant emissions, delays, and stops caused by more stable speed change and fewer stops at the intersection [38].

To our knowledge, the first approach focuses on vehicle trajectory planning [39, 40]. He et al. [32] proposed a speed optimization model to give ecodriving suggestions considering queues on a signalized arterial. Wan et al. [22] developed a speed advisory model (SAM) based on a given signal timing plan. Then, an analytical driving strategy is obtained to minimize fuel consumption. The results indicated that the SAM reduces fuel consumption and benefitted human-driven vehicles (HDVs), and the platoon fuel consumption decreased with the increase of CAVs' penetration rates. Zhao et al. [25] designed an ecological driving strategy to coordinate the platoon mixed with CAVs and HDVs. A model predictive control is proposed to save platoons' fuel consumption with a fixed-time traffic signal. The results showed that the driving strategy could further smooth out the trajectory and save fuel consumption. Therefore, these studies mainly focus on optimizing CAVs trajectories based on a preset traffic signals.

The second method optimizes signal timing plans by CAVs data [41, 42]. Goodall et al. [35] optimized traffic signal with a predictive microscopic simulation algorithm (PMSA). The connected vehicles (CVs) data, including locations and speeds, were used to predict future traffic conditions via the microscopic simulation method. A 15-second rolling horizon was chosen to minimize vehicles' delay, stops, and decelerations. Feng et al. [9] presented a real-time traffic adaptive signal control algorithm to minimize vehicle delay and queue length via connected vehicle (CVs) data. The simulation results indicated that the proposed algorithm reduced vehicle delay and balanced each phase's queue length. However, they did not consider optimizing the CAVs trajectories at the same time.

Therefore, to address this gap, the third approach simultaneously optimizes CAVs trajectories and traffic signals. Xu et al. [28] presented a two-level method to optimize traffic signal and speed for CAVs. The first level optimized traffic signals and CAVs arrival times to minimize travel time; the second level planned CAVs trajectories to save individual vehicles' fuel consumption. The results indicated that this method could improve transportation efficiency and fuel economy significantly. Yu et al. [29] developed mixed-integer linear programming to optimize vehicle trajectories and traffic signals at a signalized intersections. Simulation results showed that this method was superior to actuated control in vehicle's delay, intersection capacity, and CO_2 emission. Feng et al. [8] proposed a two-stage method with traffic signal optimization and vehicle trajectory planning. The optimal control theory and dynamic programming (DP) are applied to optimize vehicle trajectories and traffic signals to minimize vehicle delay and fuel consumption. Results showed that the proposed method could reduce vehicle delay and fuel consumption under different demand compared to fixed-time traffic signal control. However, these joint optimization methods only optimize the trajectory of the leading vehicle in a platoon; a car-following model is used to calculate the other vehicle's trajectory in the platoon. Ghiasi et al. [27] considered the joint optimization algorithm's computational efficiency; an analytical solution to joint CAVs trajectories and traffic signals optimization problem was proposed in their study. The numerical experiment showed that the proposed model could reduce travel delay and fuel consumption significantly.

This study proposes a two-level model for traffic signal timing and trajectories planning of multiple CAVs considering the random arrival of vehicles. The integrated optimization problem is modeled as a two-level model. Firstly, the traffic signal and arrival time for CAVs are optimized by the signal timing model to minimize the average vehicle's delay. Secondly, considering average gasoline consumption, an optimal control method is proposed to optimize trajectories for all CAVs. Finally, the proposed method is tested in a simulation experiment, and numerical studies and sensitivity analysis are carried out based on a simple two-phase intersection.

3. Methodology

3.1. Assumption. The following necessary assumptions are made to facilitate modeling and analysis.

(1) The interarrival time of all CAVs follows the shifted negative exponential distribution, which is verified at an isolated intersection [8, 21, 29]. This means CAVs arrive at the border of the control zone following a Poisson distribution.

(2) All CAVs can share information (such as location, speed, and arrival time) through V2V; hence, their arrival time can be predicted more accurately [25].

(3) All CAVs arrive at the boundary of the control zone and through the downstream intersection with the desired speed, which can refer to Ghiasi et al. [27].

(4) All CAVs cannot change lanes in the control zone; that is, only the longitudinal movement is considered [43–45].

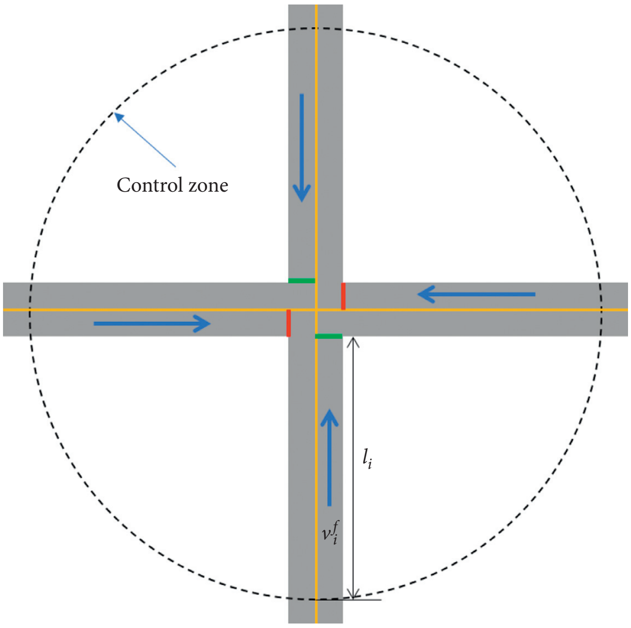


Figure 1: A simple intersection with four arms.

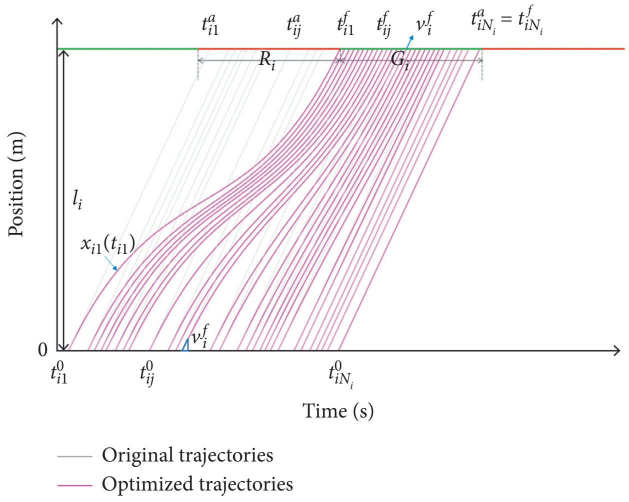


Figure 2: The vehicle trajectories planning of each arm $i.i \in \mathcal{I}$.

3.2. Problem Statement. In this study, no left-turn and right-turn are considered; only through traffic flow it is modeled, which is shown in Figure 1. There are four arms indexed by $i \in \mathcal{I} = \{1, 2, 3, 4\}$, and l_i and v_i^f are the length of the control zone and the desired speed of arm i, $i \in \mathcal{I}$, respectively. A simple two-phase signal timing plan and an arm i as an example are shown in Figure 2; the traffic signal is $\mathcal{S} = \{G_1, G_2, G_3, G_4\}$ or $\mathcal{S} = \{R_1, R_2, R_3, R_4\}$, where G_i and R_i are the effective green time and red time for arm i, $i \in \mathcal{I}$, respectively. In this study, the indexes 1, 2, 3, and 4 are defined as east, south, west, and north arm, respectively. Therefore, there have $G_1 = G_3$ and $G_2 = G_4$. Let $L = R_1 + R_2 - G_1 - G_2$ represent the lost time of a traffic signal cycle. The traffic arrival rate and the saturation flow rate of arm i are defined as λ_i and μ_i. The unsaturated traffic is considered in this study, which can be expressed as $\sum_{i \in \mathcal{I}} (\lambda_i (R_i + G_i)/\mu_i G_i) < 1$. For the convenience of the readers, the main variables used in this paper are shown in Table 1.

As shown in Figure 2, CAVs arrival at the border of the control zone is defined as $j \in \mathcal{N}_i = \{1, 2, \ldots, N_i\}$, $i \in \mathcal{I}$. Let $\mathcal{X}_i = \{x_{ij}(t_{ij})\}$ be the set of CAV trajectories at each arm i, where $x_{ij}(t_{ij})$ is the position of the j-th CAV at each arm i at time t_{ij}. $\dot{x}_{ij}(t_{ij})$ and $\ddot{x}_{ij}(t_{ij})$ are the instantaneous speed and acceleration of the j-th CAV at each arm i at time t_{ij}, respectively. Let t_{ij}^a and t_{ij}^f be the expected and optimal arrival times of the j-th CAV at the stop line of each arm i. t_{ij}^0 is the time of j-th CAV arriving at the border of the control zone at each arm i, which can be estimated accurately via advanced CAV technology [27].

Table 1: Notation of major symbols used in this paper.

Symbol	Description
Gasoline consumption	
α	The parameter of gasoline consumption rate
M	The weight of the vehicle
β_1	The parameter relevant to the energy efficiency of the engine
β_2	The parameter associated with positive acceleration
v	The vehicle's speed
a	The vehicle's acceleration
$P(t)$	The power (kW) required to drive the vehicle
Traffic signals	
$\mathscr{I}$	Set of arms at the intersection
$\mathscr{S}$	A signal timing plan
G_i	The effective green time for arm i
R_i	The effective red time for arm i
L	The lost time of a traffic signal cycle
l_i	The length of the control zone at arm i
v_i^f	The desired speed at each arm i, which is equal to free-flow speed
λ_i	The vehicle arrival rate at arm i
μ_i	The saturation flow rate at arm i
$G_i^{\min}$	The minimum green time duration for arm i
$G_i^{\max}$	The maximum green time duration for arm i
Vehicle trajectory	
$\mathscr{N}_i$	The set of CAVs arriving at the border of control zone at arm i
$\mathscr{X}_i$	The set of CAVs trajectories functions at arm i
$x_{ij}(t_{ij})$	The location of the j-th CAV at arm i at time t_{ij}
$\dot{x}_{ij}(t_{ij})$	The speed of the j-th CAV at arm i at time t_{ij}
$\ddot{x}_{ij}(t_{ij})$	The acceleration of the j-th CAV at arm i at time t_{ij}
t_{ij}^0	The time of j-th CAV arriving at the border of control zone at arm i
t_{ij}^a	The expected arrival time at the stop line of the j-th CAV at arm i
t_{ij}^f	The optimal arrival time at the stop line of the j-th CAV at arm i
τ	The delay of control and communication
s_0	The safety spacing between two consecutive CAVs
$a_{\min}$	The minimum acceleration of CAVs
$a_{\max}$	The maximum acceleration of CAVs

3.3. Model Formulation. The proposed method consists of two levels, i.e., vehicle's arrival time and traffic signal timing, and vehicle trajectories planning. The former optimizes traffic signals and vehicles' arrival time for CAVs to minimize the average vehicle's delay. The latter optimizes trajectories for all CAVs considering average gasoline consumption based on the optimal traffic signal timing plan. To better understand the proposed model, the vehicle's trajectories are optimized by giving the optimal traffic signal plan of a two-phase intersection: $\mathscr{S} = \{G_1, G_2, G_3, G_4\}$ or $\mathscr{S} = \{R_1, R_2, R_3, R_4\}$. Here, $G_1 = G_3$, $G_2 = G_4$ and $L = R_1 + R_2 - G_1 - G_2$.

3.3.1. Optimal Arrival Time. The time of CAVs (t_{ij}^0) arriving at the control zone border can be accurately estimated via the CAV technology [27]. Then, the expected arrival time of the j-th CAV arrival at the stop line of arm i can be estimated by

$$t_{ij}^a = t_{ij}^0 + \frac{l_i}{v_i^f}, \quad \forall i \in \mathscr{I}, j \in \mathscr{N}_i, \tag{1}$$

where the red signal is defined as the cycle starts is shown in Figure 2. Therefore, the number of CAVs arrival at this cycle ($N_i = |\mathscr{N}_i|$) is determined by the number of t_{ij}^a, which is determined by the arrival flow rate (λ_i).

The analysis indicates that the optimal arrival time at the stop line is determined by the expected arrival times, traffic signals, and saturation flow rate. Taking the optimal arrival time of the j-th CAV at each arm i as an example, it can be divided into the following four cases.

(a) The first CAV of a signal cycle at each arm i:

(i) If the expected arrival time of the first CAV is during the red signal period, to minimize the vehicle's delay, the first CAV's optimal arrival time is equal to the start time of the green signal in the next signal cycle.

(ii) If the expected arrival time of the first CAV is during the green signal duration, to minimize the vehicle's delay, the first CAV's optimal arrival time is equal to the expected arrival time.

(b) The other CAVs of a signal cycle at each arm i: the estimated arrival time is the sum of the optimal arrival time of the preceding CAV and saturation headway.

(i) If the estimated arrival time is shorter than the expected arrival time, the optimal arrival time is equal to the expected arrival time at the stop line.

(ii) If the estimated arrival time is not shorter than the expected arrival time, the optimal arrival time is equal to the estimated arrival time.

$$t_{ij}^{f}=\begin{cases} t_{i(j-1)}^{f}+\dfrac{1}{\mu_i}, & \text{if } t_{ij}^{a}\le t_{i(j-1)}^{f}+\dfrac{1}{\mu_i} \\ t_{ij}^{a}, & \text{if } t_{ij}^{a}> t_{i(j-1)}^{f}+\dfrac{1}{\mu_i} \end{cases},\quad \forall i\in\mathcal{I},\ j\in\mathcal{N}_i\setminus\{1\}. \tag{3}$$

3.3.2. Objective Function. (1) Vehicle's Delay Function. The travel delay of each CAV is defined as the difference between the actual and the free travel time. The free and actual travel time can be determined by (1) and (3), respectively. As a result, the vehicle's delay function for arm i is formulated as

$$\mathcal{D}_i\left(\mathcal{S},\mathcal{X}_i\right)=\frac{1}{N_i}\sum_{j\in\mathcal{N}_i}\left(t_{ij}^{f}-t_{ij}^{a}-\frac{l_i}{v_i^{f}}\right),\quad \forall i\in\mathcal{I}, \tag{4}$$

where $\mathcal{D}_i$ is the average vehicle's delay for each arm i.

Therefore, the average vehicle's delay for this intersection is formulated as

$$\mathcal{D}\left(\mathcal{S},\mathcal{X}\right)=\frac{1}{\sum_{i\in\mathcal{I}}N_i}\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{N}_i}\left(t_{ij}^{f}-t_{ij}^{a}-\frac{l_i}{v_i^{f}}\right). \tag{5}$$

(2) Gasoline Consumption Function. Gasoline consumption is a function of instantaneous speed and acceleration of vehicle [46–48], which is formulated as

$$F\left(v,a;t\right)=\begin{cases} \alpha+\beta_1 R_T(t)v(t)+\max\left\{0,\dfrac{\beta_2 Ma^2(t)v(t)}{1000}\right\}, & \text{if } R_T(t)>0, \\ \alpha, & \text{if } R_T(t)\le 0, \end{cases} \tag{6}$$

where α represents constant idle fuel rate (ml/s), M represents the weight of the vehicle (kg), β_1 and β_2 represent the efficiency parameters, a and v represent instantaneous acceleration and speed of a vehicle, respectively, and $R_T(t)$ represents total "tractive" force required to drive the vehicle, which is defined as

$$R_T(t)=b_1+b_2v(t)+b_3v^2(t)+\frac{Ma(t)}{1000}+9.81\times10^{-5}MG. \tag{7}$$

where b_1, b_2, and b_3 represent rolling, engine, and aerodynamic drag, respectively; G is percent grade. Referring to Akcelik [47], the calibrated parameters in (6) and (7) are $M = 1600\,\text{kg}$, $G = 0$, $\alpha = 0.666\ \text{ml/kJ}$, $\beta_1 = 0.0717\ \text{ml/kJ}$, $\beta_2 = 0.0344\ \text{ml/}\left(\text{kJ}\cdot\text{m/s}^2\right)$, $b_1 = 0.269\,\text{kN}$, $b_2 = 0.0171\,\text{kN}\ (\text{m/ s}^2)$, $b_3 = 0.000672$.

The average gasoline consumption function for arm i is defined as

$$\mathcal{G}_i\left(\mathcal{S},\mathcal{X}_i\right)=\frac{1}{N_i}\sum_{j\in\mathcal{N}_i}\int_{t_{ij}^{0}}^{t_{ij}^{f}}F\left(\dot{x}_{ij}(t),\ \ddot{x}_{ij}(t);t\right)\mathrm{d}t,\quad \forall i\in\mathcal{I}, \tag{8}$$

where $\mathcal{G}_i$ is the average gasoline consumption for arm i.

Therefore, the average gasoline consumption for this intersection is formulated as

$$\mathcal{G}\left(\mathcal{S},\mathcal{X}\right)=\frac{1}{\sum_{i\in\mathcal{I}}N_i}\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{N}_i}\int_{t_{ij}^{0}}^{t_{ij}^{f}}F\left(\dot{x}_{ij}(t),\ \ddot{x}_{ij}(t);t\right)\mathrm{d}t. \tag{9}$$

3.3.3. Constrain Conditions

(1) Traffic Signals Constrain. The green time duration constraints: the green time duration of each arm i must be between the minimum and maximum green time duration.

$$G_i^{\min}\le G_i\le G_i^{\max},\quad \forall i\in\mathcal{I}, \tag{10}$$

where $G_i^{\min}$ and $G_i^{\max}$ are the minimum and maximum green time duration for arm i, respectively.

The Signal Cycle Constraint. The sum-up of effective red time duration for all phases must equal the sum up of effective green time duration and constant lost time.

$$R_1+R_2=G_1+G_2+L. \tag{11}$$

The Unsaturated Traffic Flow Constraint. The maximum number of the departure CAVs must not be smaller than the number of the arrival CAVs for each arm i.

$$\lambda_i\left(R_i+G_i\right)\le\mu_i G_i,\quad \forall i\in\mathcal{I}. \tag{12}$$

(2) Vehicle Trajectories Constrain. Dynamic state constraint: at arm i, the position, velocity, and acceleration of the j-th CAV at any time should satisfy the following dynamic equations.

$$\dot{x}_{ij}(t)=\frac{\mathrm{d}x_{ij}(t)}{\mathrm{d}t},\quad \forall t\in\left[t_{ij}^{0},t_{ij}^{f}\right], i\in\mathcal{I}, j\in\mathcal{N}_i,\quad \ddot{x}_{ij}(t)=\frac{\mathrm{d}\dot{x}_{ij}(t)}{\mathrm{d}t},\quad \forall t\in\left[t_{ij}^{0},t_{ij}^{f}\right], i\in\mathcal{I}, j\in\mathcal{N}_i. \tag{13}$$

Initial Boundary Constraint. At arm i, the position, velocity, and acceleration of the j-th CAV at start time are given by the assumptions [27].

$$x_{ij}\left(t^0_{ij}\right) = 0, \quad \forall i \in \mathcal{I}, j \in \mathcal{N}_i,$$
$$\dot{x}_{ij}\left(t^0_{ij}\right) = v^f_i, \quad \forall i \in \mathcal{I}, j \in \mathcal{N}_i, \left(t^0_{ij}\right) = 0, \quad \forall i \in \mathcal{I}, j \in \mathcal{N}_i.$$
$$\ddot{x}_{ij} \tag{14}$$

Final Boundary Constraint. At arm i, the position, velocity, and acceleration of the j-th CAV at end time are given by the assumptions [27].

$$x_{ij}\left(t^f_{ij}\right) = l_i, \quad \forall i \in \mathcal{I}, j \in \mathcal{N}_i,$$
$$\dot{x}_{ij}\left(t^f_{ij}\right) = v^f_i, \quad \forall i \in \mathcal{I}, j \in \mathcal{N}_i, \left(t^f_{ij}\right) = 0, \quad \forall i \in \mathcal{I}, j \in \mathcal{N}_i.$$
$$\ddot{x}_{ij} \tag{15}$$

Consecutive Vehicle Position Constraint. The adjacent CAVs must meet the specific safety headway because of control and communication delay. The headway between vehicle $(j-1)$'s location with a control and communication delay τ ago $x_{i(j-1)}(t-\tau)$ and vehicle j's location, $x_{ij}(t)$ is no less than s_0 in time interval $[t^0_{i(j-1)}, t^f_{ij}]$.

$$x_{i(j-1)}(t-\tau) - x_{ij}(t) \geq L + s_0, \quad \forall i \in \mathcal{I}, n \in \mathcal{N}_i \setminus \{1\}, t_{ij} \in \left[t^0_{i(j-1)}, t^f_{ij}\right], \tag{16}$$

where τ is control and communication delay, s_0 is the safety spacing between two adjacent CAVs, and$L$ is the length of CAVs.

Speed Constraint. The speed of all CAVs cannot go beyond the free speed limit.

$$0 \leq \dot{x}_{ij}(t) \leq v^f_i, \quad \forall t \in \left[t^0_{ij}, t^f_{ij}\right], i \in \mathcal{I}, j \in \mathcal{N}_i. \tag{17}$$

Acceleration Constraint. The acceleration of all CAVs must be between the minimum and maximum acceleration.

$$a_{\min} \leq \ddot{x}_{ij}(t) \leq a_{\max}, \quad \forall t \in \left[t^0_{ij}, t^f_{ij}\right], i \in \mathcal{I}, j \in \mathcal{N}_i, \tag{18}$$

where $a_{\min}$ and $a_{\max}$ are the minimum and maximum acceleration, respectively.

3.4. Solution Method. In this study, a dynamic programming (DP) algorithm and the GPOPS are adopted to solve the traffic signal timing problem and multiple vehicle trajectories planning problem, respectively.

3.4.1. Dynamic Programming. Many DP-based traffic signal timing methods have been developed [8, 9, 49]. In the DP algorithm, state variables and decision variables are the key parameters. Equations (19)–(20) illustrate the relationship between the two parameters; see more details in [49].

$$s_p = s_{p-1} + h\left(x_p\right), \tag{19}$$

$$h\left(x_p\right) = \begin{cases} 0, & \text{if } x_p = 0, \\ x_p + r_p, & \text{otherwise,} \end{cases} \tag{20}$$

where s_p is the total number of time intervals from the beginning stage to the end of stage p and x_p and r_p are the green and the clearance time intervals of the stage p.

When the state variable s_p is given, the feasible set of decision variables can be calculated by

$$X_p\left(s_p\right) = \begin{cases} 0, & if\, s_p - r_p < x_{\min}, \\ \{x_{\min}, x_{\min}+1, \ldots, x_{\max}\}, & if\, s_p - r_p \geq x_{\min} \text{ and } T - s_{p-1} - r_p > x_{\max}, \\ \left\{x_{\min}, x_{\min}+1, \ldots, T - s_{p-1} - r_p\right\} & if\, T - s_{p-1} - r_p \leq x_{\max}. \end{cases} \tag{21}$$

After determining $X_p(s_p)$, DP is adopted to search for the optimal decision variables x_p. The DP algorithm consists of two recursions; the first recursion obtains the optimal objective function in every time interval; the second recursion searches the decision variables corresponding to the optimal objective.

3.5. Forward Recursion

(i) Step 1: Set initial stage $p = 1$, state variable $s_{p-1} = 0$, and value function $v_p(s_{p-1}) = 0$.

(ii) Step 2: For $s_p = 1, 2, \ldots, T\{$

$v_p(s_p) = \min\left\{f_p(s_p, x_p) + v_{p-1}(s_{p-1}) | x_p \in \mathbf{X}_p(s_p)\right\}$

$x^*_p(s_p) =$
$\text{argmin}_{x_p}\left\{f_p(s_p, x_p) + v_{p-1}(s_{p-1}) | x_p \in \mathbf{X}_p(s_p)\right\}$

Record $x^*_p(s_p)$ and $v_p(s_p)$ as the optimal solution and value function$\}$.

(iii) Step 3: If $(p < |P|)$, let$p = p + 1$, and go to Step 2.
Else if $(v_{p-k}(T) = v_p(T))$ for all $k \leq |P| - 1$, STOP.
Else $p = p + 1$, go to Step 2.

The first recursion starts with stage 1 and the cumulative value function as 0. For each stage, the DP searches the optimal solution $x^*_p(s_p)$ with a given state variable s_p. The objective function $f_p(s_p, x_p)$ is determined by the expected arrival time (1) of all CAVs. The stop criteria for the first recursion are derived from Sen and Head [49]. Besides, the number of phases $|P|$ is 2 in this study, which contains the east-west phase and the north-south phase.

3.6. Backward Recursion. After optimal value function is determined, the optimal decision $x^*_p(s_p)$ of each stage can be retrieved in the second recursion as follows.

(i) **Step 1:** Set the optimal stages as J, and the optimal state variable $s^*_{J-1} = T$.

(ii) **Step 2:** For $p = J-1, J-2, \ldots, 1\{$

Finding $x^*_p(s^*_p)$ from the records of **Forward recursion.**

If $(j > 1)$, $s^*_{p-1} = s^*_p - h_p(x^*_p(s^*_p))\}$.

3.6.1. General Pseudospectral Optimal Control Method. As an optimal control problem, the vehicle trajectories planning can be handled numerically by GPOPS [30], which is widely used in vehicle trajectory optimization [25, 32, 33]. Therefore, the GPOPS is used to solve the optimal control problem for multiple CAVs trajectory planning.

3.6.2. Solution Algorithm. In summary, the two-level optimization algorithm is as follows. (i.e., Algorithm 1).

4. Numerical Studies

4.1. Simulation Settings. The simulation duration of every scenario with a different traffic volume is 900 seconds. Every scenario is repeated five times with different random seeds. Besides, vehicle arrival conforms to the Poisson distribution [8, 21, 29].

In signal optimization, a four-arm and two phases of a cycle are selected. The time planning horizon is $T_p = 50\,\text{s}$. The minimum and maximum green time are $G_i^{\min} = 15\,\text{s}$ and $G_i^{\max} = 30\,\text{s}$, respectively. The lost time of each phase $(L/2) = 1\,\text{s}$. The length of the control zone at each arm $l_i = 300\,\text{m}$, and the free flow and the desired speed at each arm $v_i^f = 15\,\text{m/s}$. The saturation flow rate of each arm $\mu_i = 1\,\text{veh/s}$, which must be less than $1/(s_0/v_i^f) = 3\,\text{veh/s}$ in this study.

In the vehicle trajectories planning, the delay of control and communication $\tau = 0.1\,\text{s}$. The safety spacing between two consecutive CAVs $s_0 = 5\,\text{m}$. The length of CAVs $L = 5\,\text{m}$. The minimum and maximum acceleration are $a_{\min} = -6\,\text{m/s}^2$ and $a_{\max} = 3\,\text{m/s}^2$, respectively.

4.2. Results and Discussions. The two-level integrated optimization model, denoted as "IO", is compared with Signal-fixed. Three volume levels, namely, 600, 800, and 1200 vph, are created in this study [50]. The demands in the two approaches (i.e., arm 1 and 3, arm 2 and 4) are set to be the same. To consider the difference in traffic between the two directions, we designed four scenarios, including two balanced and two unbalanced flows. In the "IO" control, vehicle trajectories are optimized by GPOPS [30], and the DP algorithm optimizes the signal timing plan in different scenarios. In the "Signal-fixed" control, vehicle trajectories are optimized by GPOPS [30], and the signal timing plan is optimized by Synchro [51] in different scenarios. Specifically, the signal parameters setup is the same as "IO" (e.g., the lost time of each phase, the saturation flow rate, and the minimum and maximum green time). The average vehicle's delay and gasoline consumption of 4 scenarios with different traffic demands are shown in Table 2. Besides, all CAVs trajectories and traffic signal plans can be obtained. Figure 3 shows vehicle trajectories for 4 scenarios with different demand.

As shown in Table 2, there are four scenarios, namely, 1200/1200, 1200/800, 800/800, and 800/600 vph. The simulation results show a significant decrease in the average vehicle's delay and gasoline consumption when IO control is applied. Compared with the Signal-fixed, the reduced average vehicle's delay with four scenarios are 26.91%, 15.57%, 24.17%, and 21.77%, and the reduced gasoline consumption with four scenarios are 10.38%, 5.30%, 8.50%, and 7.15%. In other words, the proposed integrated optimization method can averagely improve the transportation efficiency by 21.77% and decrease gasoline consumption by 7.83%, compared with Signal-fixed control in these studied scenarios, respectively.

Figure 3 shows that all CAVs pass through the intersection at free speed without stopping. Therefore, no CAVs are queuing at the stop line of the intersection. Furthermore, this method eliminates the loss of green start-up time compared with no trajectory optimization, and more vehicles can pass through the intersection in the same green interval. Besides, compared with Signal-fixed control, IO control has a smaller vehicle delay and gasoline consumption. This indicates that the integrated optimization method can better consider traffic signal and vehicle trajectories optimization, thus further reducing the average vehicle's delay and gasoline consumption, compared with Signal-fixed control. In addition, the minimum green time duration is considered in this study. Therefore, a part of the green time duration of the phase is wasted in Figure 3.

4.3. Sensitivity Analysis. In this study, the minimum green time $(G_i^{\min})$ and free-flow speed (v_i^f) are the most critical parameters. Therefore, we have carried on the analysis and the discussion of these two parameters.

4.3.1. Minimum Green Time. Minimum green time is to ensure the safety of drivers and pedestrians. A minimum green time that is too long may result in increased delay; one that is too short may violate pedestrian needs. Therefore, different geometric shapes of intersections can set different minimum green time. To avoid the influence of other parameters, scenario one (1200/1200 vph) is selected as a sensitivity analysis of the minimum green time. In the sensitivity analysis, $G_i^{\min}$ varies from 10 s to 20 s with an

```
Initialize:
Set the total simulation time as $T$, the time planning horizon as $T_p$, the current time as $T_c = 0, L, \lambda_i, \mu_i, G_i^{\min}, G_i^{\max}, l_i, v_i^f$ in arm $i, \forall i \in \mathcal{I}$.
Simulate the arrival times of CAVs at arm $i, \forall i \in \mathcal{I}$.
Iterate:
While $T_c + T_p \leq T$ do
    Get the arrival times ($t_{ij}^0$, $\forall i \in \mathcal{I}, j \in \mathcal{N}_i$) of CAVs in time planning horizon $[T_c, T_c + T_p]$.
    Calculate $t_{ij}^a$, $\forall i \in \mathcal{I}, j \in \mathcal{N}_i$ based on equation (1).
    Optimize the traffic signal timing plan by DP algorithm.
    For Each signal cycle do
        Obtain signal time plan $\mathcal{S}$.
        For $i = 1 \longrightarrow 4$ do
            Obtain $\mathcal{N}_i, t_{ij}^0, t_{ij}^a$, $\forall i \in \mathcal{I}, j \in \mathcal{N}_i$.
            Calculate $t_{ij}^f$, $\forall i \in \mathcal{I}, j \in \mathcal{N}_i$.
            Optimize the $j$-th CAV trajectory by GPOPS.
            Save the vehicle trajectories $\mathcal{X}_i$ for the $i$ arm.
            Calculate $\mathcal{G}_i$ and $\mathcal{D}_i$.
        End
        Calculate $\mathcal{D}$ and $\mathcal{G}$.
    End
    Save vehicle trajectories, signal timing plan, gasoline consumption, and average delay at the current time planning horizon $[T_c, T_c + T_p]$.
    $T_c = T_c + T_p$
End
Output:
Output vehicle trajectories, signal timing plan, gasoline consumption, and average delay at the total time planning horizon $[0, T]$.
```

Algorithm 1

increment of 1 s. The sensitivity analysis result is shown in Figure 4.

As shown in Figure 4, the sensitivity analysis result shows that a shorter minimum green time results in a significantly less average vehicle's delay and gasoline consumption under IO control. In the unsaturated traffic flow, a shorter minimum green time can ensure that CAVs pass through intersections faster, resulting in less travel time, deceleration, and acceleration. This is because a shorter minimum green time helps avoid the waste of green time caused by the random arrival of vehicles, especially in low traffic flow rates. As a result, there are a smaller average vehicle's delay and lower gasoline consumption.

4.3.2. Free-Flow Speed. The free-flow speeds influence CAVs arrival time, which is an essential parameter for traffic signal optimization and trajectories planning of this study. Scenario no.1 (1200/1200 vph) is selected as a sensitivity analysis of the free-flow speeds. In the sensitivity analysis, v_i^f is from 10 m/s to 20 m/s in steps of 1 m/s. The sensitivity analysis result is shown in Figure 5.

The sensitivity analysis (Figure 5) shows that the average vehicle's delay decreases with free-flow speed. This indicates that a more significant free speed resulting in shorter travel times of CAVs would lead to smaller vehicle delays. However, Figure 5 indicates the average gasoline consumption decreases with free-flow speed (10–13 m/s) before reaching the lowest point when the free speed is 13 m/s and then starts to increase. This suggests an optimal free-flow speed to minimize the average gasoline consumption, and the optimal free-flow speed is 13 m/s in this scenario.

5. Conclusions and Future Work

This study developed a two-level model for traffic signal timing and trajectories planning of multiple connected automated vehicles considering the random arrival of vehicles. Based on the numerical experiments, the following conclusions can be drawn:

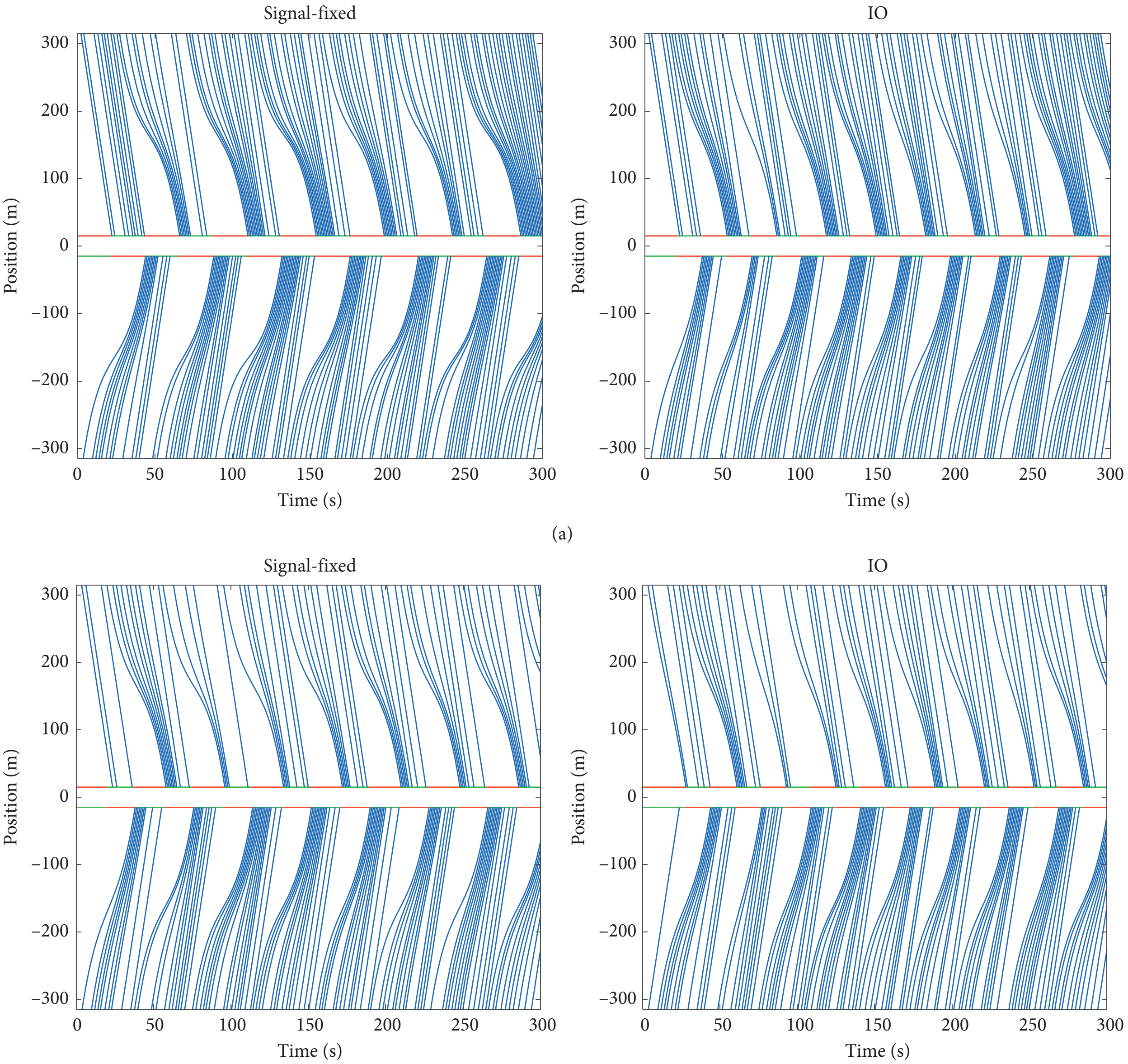
(1) Compared with the Signal-fixed, the reduced average vehicle's delays with four scenarios are 26.91%, 15.57%, 24.17%, and 21.77%, and the reduced gasoline consumption with four scenarios are 10.38%, 5.30%, 8.50%, and 7.15%.

(2) The proposed two-level model could reduce both vehicle's delay and gasoline consumption by 26.91% and 10.38%, compared with Signal-fixed control in these studied scenarios, respectively.

(3) Sensitivity analysis suggests that the minimum green time and free speed have a significant impact on the two-level model's performance.

(4) A shorter minimum green time results in a significantly less average vehicle's delay and gasoline consumption. The optimal free-flow speed is 13 m/s in the study scenario.

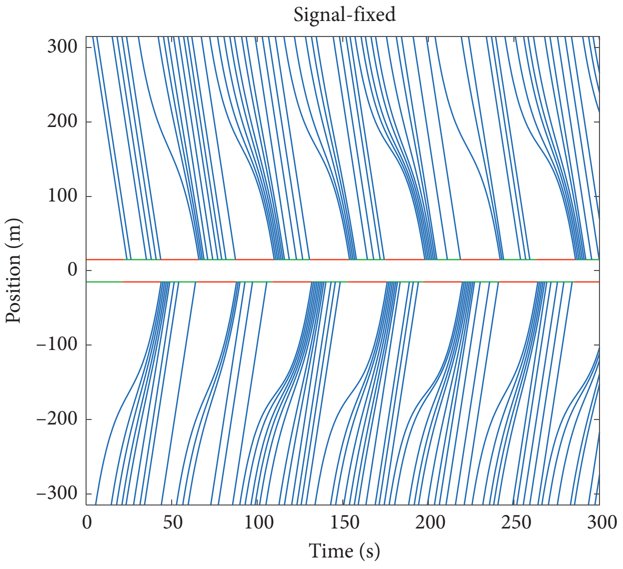
In the current work, this work applied the proposed model to a single intersection, similar to vehicle merging



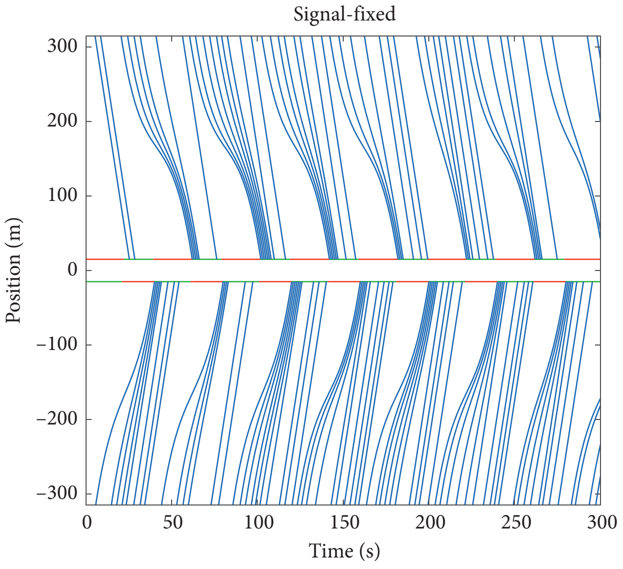
(a)

(b)

Figure 3: Continued.



(c)



(d)

Figure 3: Trajectories of CAVs in arm 1 and 2 as an example. (a) 1200/1200 vph. (b) 1200/800 vph. (c) 800/800 vph. (d) 800/600 vph.

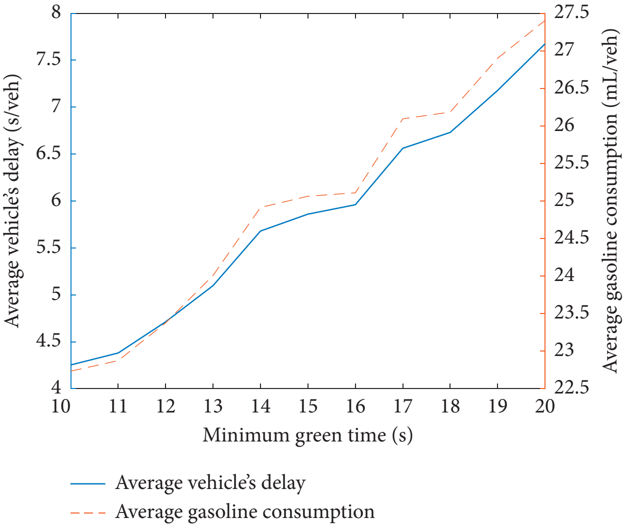


Figure 4: Sensitivity analysis on minimum green time.

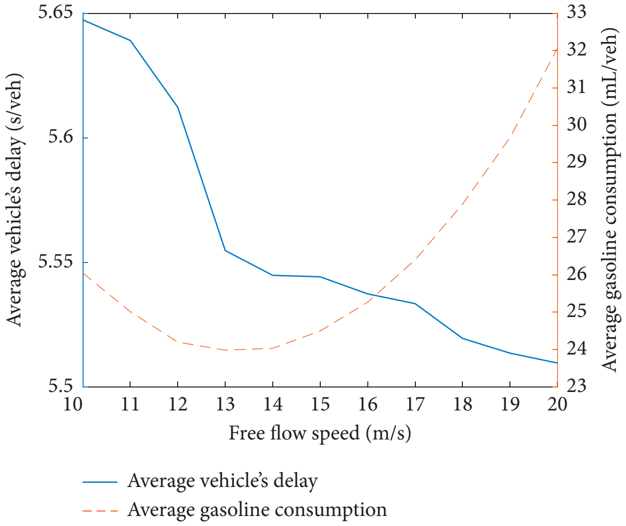


Figure 5: Sensitivity analysis on free-flow speed.

Table 2: The average vehicle's delay and gasoline consumption in different scenarios.

λ_1 (vph)	λ_2 (vph)	Delay (s/veh)			Gasoline consumption (mL/veh)		
		IO	Signal-fixed	Decrease	IO	Signal-fixed	Decrease
1200	1200	5.8598	8.0169	−26.91%	25.0633	27.9663	−10.38%
1200	800	5.4404	6.4440	−15.57%	24.3265	25.6891	−5.30%
800	800	5.0441	6.6516	−24.17%	23.2960	25.4607	−8.50%
800	600	4.9859	6.2673	−20.45%	23.2806	25.0732	−7.15%
Average	5.3326	6.8450	−21.77%	23.9916	26.0473	−7.83%	

behavior [50, 52, 53]. We will improve the proposed model and apply it to multiple intersections or a traffic network in the next step.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work are supported by the National Natural Science Foundation of China (no. 52002339), the Science and Technology Program of Sichuan Province (2021YJ0535), and the Fundamental Research Funds for the Central Universities (2682021CX058).

References

[1] W. Li and X. Ban, "Connected vehicles based traffic signal timing optimization," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 12, pp. 4354–4366, 2019.

[2] Z. Yao, B. Zhao, T. Yuan, H. Jiang, and Y. Jiang, "Reducing gasoline consumption in mixed connected automated vehicles environment: a joint optimization framework for traffic signals and vehicle trajectory," *Journal of Cleaner Production*, vol. 265, Article ID 121836, 2020.

[3] B. Schneider, *New Study of Global Traffic Reveals that Traffic is Bad*, City Lab, Los Angeles, CA, USA, 2018.

[4] Z. Li, P. Liu, C. Xu, H. Duan, and W. Wang, "Reinforcement learning-based variable speed limit control strategy to reduce traffic congestion at freeway recurrent bottlenecks," *IEEE Trans. Intell. Transport. Syst.*vol. 18, no. 11, pp. 3204–3217, 2017.

[5] Q. Wan, G. Peng, Z. Li, and F. H. T. Inomata, "Spatiotemporal trajectory characteristic analysis for traffic state transition prediction near expressway merge bottleneck," *Transportation Research Part C: Emerging Technologies*, vol. 117, Article ID 102682, 2020.

[6] A. H. F. Chow, S. Li, and R. Zhong, "Multi-objective optimal control formulations for bus service reliability with traffic signals," *Transportation Research Part B: Methodological*, vol. 103, pp. 248–268, 2017.

[7] F. V Webster, *Traffic Signal Settings*, Department of Scientific and Industrial Research, London, UK, 1958.

[8] Y. Feng, C. Yu, and H. X. Liu, "Spatiotemporal intersection control in a connected and automated vehicle environment," *Transportation Research Part C: Emerging Technologies*, vol. 89, pp. 364–383, 2018.

[9] Y. Feng, K. L. Head, S. Khoshmagham, and M. Zamanipour, "A real-time adaptive signal control in a connected vehicle environment," *Transportation Research Part C: Emerging Technologies*, vol. 55, pp. 460–473, 2015.

[10] Y. Feng, M. Zamanipour, K. L. Head, and S. Khoshmagham, "Connected vehicle-based adaptive signal control and applications," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2558, no. 1, pp. 11–19, 2016.

[11] B. Beak, K. L. Head, and Y. Feng, "Adaptive coordination based on connected vehicle technology," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2619, no. 1, pp. 1–12, 2017.

[12] J. Lee, B. Park, I. Yun, and I. Yun, "Cumulative travel-time responsive real-time intersection control algorithm in the connected vehicle environment," *Journal of Transportation Engineering*, vol. 139, no. 10, pp. 1020–1029, 2013.

[13] C. Priemer and B. Friedrich, "A decentralized adaptive traffic signal control using v2i communication data," in *Proceedings of the 2009 12th International IEEE Conference On Intelligent Transportation Systems*, pp. 765–770, St. Louis, MO, USA, October 2009.

[14] Z. Yao, H. Jiang, Y. Cheng, Y. Jiang, and B. Ran, "Integrated schedule and trajectory optimization for connected automated vehicles in a conflict zone," *IEEE Transactions on Intelligent Transportation Systems*, vol. 40, pp. 1–11, 2020.

[15] F. Zhu and S. V. Ukkusuri, "A linear programming formulation for autonomous intersection control within a dynamic traffic assignment and connected vehicle environment," *Transportation Research Part C: Emerging Technologies*, vol. 55, pp. 363–378, 2015.

[16] X. Zeng, X. Sun, Y. Zhang, and L. Quadrifoglio, "Person-based adaptive priority signal control with connected-vehicle information," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2487, no. 1, pp. 78–87, 2015.

[17] E. Bagheri, B. Mehran, B. Hellinga, and Bruce, "Real-time estimation of saturation flow rates for dynamic traffic signal control using connected-vehicle data," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2487, no. 1, pp. 69–77, 2015.

[18] K. Tiaprasert, Y. Zhang, X. B. Wang, and X. Zeng, "Queue length estimation using connected vehicle technology for adaptive signal control," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 4, pp. 2129–2140, 2015.

[19] K. Yang, S. I. Guler, and M. Menendez, "A transit signal priority algorithm under connected vehicle environment," in *Proceedings of the 2015 IEEE 18th International Conference on Intelligent Transportation Systems*, pp. 66–70, Gran Canaria, Spain, October 2015.

[20] Y. Wang, W. Ma, W. Yin, and X. Yang, "Implementation and testing of cooperative bus priority system in connected vehicle

environment," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2424, no. 1, pp. 48–57, 2014.

[21] H. Jiang, J. Hu, S. An, M. Wang, and B. B. Park, "Eco approaching at an isolated signalized intersection under partially connected and automated vehicles environment," *Transportation Research Part C: Emerging Technologies*, vol. 79, pp. 290–307, 2017.

[22] N. Wan, A. Vahidi, and A. Luckow, "Optimal speed advisory for connected vehicles in arterial roads and the impact on mixed traffic," *Transportation Research Part C: Emerging Technologies*, vol. 69, pp. 548–563, 2016.

[23] Y. Wei, C. Avcı, J. Liu et al., "Dynamic programming-based multi-vehicle longitudinal trajectory optimization with simplified car following models," *Transportation Research Part B: Methodological*, vol. 106, pp. 102–129, 2017.

[24] A. Omidvar, M. Pourmehrab, P. Emami et al., "Deployment and testing of optimized autonomous and connected vehicle trajectories at a closed-course signalized intersection," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2672, no. 19, pp. 45–54, Article ID 036119811878279, 2018.

[25] W. Zhao, D. Ngoduy, S. Shepherd, R. Liu, and M. Papageorgiou, "A platoon based cooperative eco-driving model for mixed automated and human-driven vehicles at a signalised intersection," *Transportation Research Part C: Emerging Technologies*, vol. 95, pp. 802–821, 2018.

[26] X. He and X. Wu, "Eco-driving advisory strategies for a platoon of mixed gasoline and electric vehicles in a connected vehicle system," *Transportation Research Part D: Transport and Environment*, vol. 63, pp. 907–922, 2018.

[27] A. Ghiasi, X. Li, Z. Huang, and X. Qu, "A joint trajectory and signal optimization model for connected automated vehicles," in *Proceedings of the Transportation Research Board 98th Annual Meeting*, pp. 1–10, Washington, DC, USA, January 2019.

[28] B. Xu, X. J. Ban, Y. Bian et al., "Cooperative method of traffic signal optimization and speed control of connected vehicles at isolated intersections," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 4, pp. 1390–1403, 2019.

[29] C. Yu, Y. Feng, H. X. Liu, W. Ma, and X. Yang, "Integrated optimization of traffic signals and vehicle trajectories at isolated urban intersections," *Transportation Research Part B: Methodological*, vol. 112, pp. 89–112, 2018.

[30] A. V. Rao, D. A. Benson, C. Darby et al., "Acm transactions on mathematical software Algorithm 902," *GPOPS , A MATLAB Software for Solving Multiple-Phase Optimal Control Problems Using the Gauss Pseudospectral Method*, vol. 37, no. 2, pp. 1–39, 2011.

[31] L. Li and X. Li, "Parsimonious trajectory design of connected automated traffic," *Transportation Research Part B: Methodological*, vol. 119, pp. 1–21, 2019.

[32] X. He, H. X. Liu, and X. Liu, "Optimal vehicle speed trajectory on a signalized arterial with consideration of queue," *Transportation Research Part C: Emerging Technologies*, vol. 61, pp. 106–120, 2015.

[33] X. Wu, X. He, G. Yu, A. Harmandayan, and Y. Wang, "Energy-optimal speed control for electric vehicles on signalized arterials," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 5, pp. 2786–2796, 2015.

[34] X. T. Yang, K. Huang, Z. Zhang, Z. A. Zhang, and F. Lin, "Eco-driving system for connected automated vehicles: multi-objective trajectory optimization," *IEEE Transactions on Intelligent Transportation Systems*, vol. 1, no. 1–13, 2021.

[35] N. J. Goodall, B. L. Smith, B. Park, and Park, "Traffic signal control with connected vehicles," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2381, no. 1, pp. 65–72, 2013.

[36] X. Li, A. Ghiasi, Z. Xu, and X. Qu, "A piecewise trajectory optimization model for connected automated vehicles: exact optimization algorithm and queue propagation analysis," *Transportation Research Part B: Methodological*, vol. 118, pp. 429–456, 2018.

[37] R. Niroumand, M. Tajalli, L. Hajibabai, and A. Hajbabaie, "Joint optimization of vehicle-group trajectory and signal timing: introducing the white phase for mixed-autonomy traffic stream," *Transportation Research Part C Emerging Technologies*, vol. 116, Article ID 102659, 2020.

[38] X. Chang, J. Rong, H. Li, Y. Wu, and X. Zhao, "Impact of connected vehicle environment on driving performance: a case of an extra-long tunnel scenario," *IET Intelligent Transport Systems*, vol. 15, no. 3, pp. 423–431, 2021.

[39] G. Li, S. Fang, J. Ma, and J. Cheng, "Modeling merging acceleration and deceleration behavior based on gradient-boosting decision tree," *Journal of Transportation Engineering, Part A: Systems*, vol. 146, no. 7, Article ID 05020005, 2020.

[40] G. Li, Z. Yang, Q. Yu, J. Ma, and S. Fang, "Characterizing heterogeneity among merging positions: comparison study between random parameter and latent class Accelerated hazard model," *Journal of Transportation Engineering, Part A: Systems*, vol. 147, no. 6, 2021.

[41] Z. Yao, L. Shen, R. Liu, Y. Jiang, and X. Yang, "A dynamic predictive traffic signal control framework in a cross-sectional vehicle infrastructure integration environment," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 4, pp. 1455–1466, 2020.

[42] Z. Yao, Y. Jiang, B. Zhao, X. Luo, and B. Peng, "A dynamic optimization method for adaptive signal control in a connected vehicle environment," *Journal of Intelligent Transportation Systems*, vol. 24, no. 2, pp. 184–200, 2020.

[43] Z. Yao, R. Hu, Y. Jiang, and T. Xu, "Stability and safety evaluation of mixed traffic flow with connected automated vehicles on expressways," *Journal of Safety Research*, vol. 75, pp. 262–274, 2020.

[44] Z. Yao, R. Hu, Y. Wang, Y. Jiang, B. Ran, and Y. Chen, "Stability analysis and the fundamental diagram for mixed connected automated and human-driven vehicles," *Physica A: Statistical Mechanics and Its Applications*, vol. 533, Article ID 121931, 2019.

[45] Z. Yao, T. Xu, Y. Jiang, and R. Hu, "Linear stability analysis of heterogeneous traffic flow considering degradations of connected automated vehicles and reaction time," *Physica A: Statistical Mechanics and Its Applications*, vol. 561, Article ID 125218, 2021.

[46] J. N. Hooker, "Optimal driving for single-vehicle fuel economy," *Transportation Research Part A: General*, vol. 22, no. 3, pp. 183–201, 1988.

[47] R. Akcelik, "Efficiency and drag in the power-based model of fuel consumption," *Transportation Research Part B: Methodological*, vol. 23, no. 5, pp. 376–385, 1989.

[48] K. Ahn, *Microscopic Fuel Consumption and Emission Modeling*, " Virginia Polytechnic Institute and State University, Blacksburg, VA, USA, 1998.

[49] S. Sen and K. L. Head, "Controlled optimization of phases at an intersection," *Transportation Science*, vol. 31, no. 1, pp. 5–17, 1997.

[50] G. Li and J. Cheng, "Exploring the effects of traffic density on merging behavior," *IEEE Access*, vol. 7, pp. 51608–51619, 2019.

[51] D. Husch and A. John, *Synchro 6: Traffic Signal Software, User Guide*, Trafficware, Albany, CA, USA, 2003.

[52] G. Li, J. Ma, and Q. Shen, "Modeling of merging decision during execution period based on random forest," *Journal of Advanced Transportation*, vol. 2021, Article ID 6654096, 11 pages, 2021.

[53] G. Li, Y. Pan, Z. Yang, and J. Ma, "Modeling vehicle merging position selection behaviors based on a finite mixture of linear regression models," *IEEE Access*, vol. 7, pp. 158445–158458, 2019.

Hindawi
Journal of Advanced Transportation
Volume 2021, Article ID 5585131, 14 pages
https://doi.org/10.1155/2021/5585131

Research Article

Map Matching for Fixed Sensor Data Based on Utility Theory

Kangkang He,[1] Qi Cao,[1] Gang Ren ,[1] Dawei Li,[1] and Shuichao Zhang[2]

[1]*School of Transportation, Southeast University, Nanjing 211189, China*
[2]*School of Civil and Transportation Engineering, Ningbo University of Technology, Ningbo 315211, China*

Correspondence should be addressed to Gang Ren; rengang@seu.edu.cn

Received 14 January 2021; Revised 20 February 2021; Accepted 5 March 2021; Published 27 March 2021

Academic Editor: Hai Ying Li

Map matching can provide useful traffic information by aligning the observed trajectories of vehicles with the road network on a digital map. It has an essential role in many advanced intelligent traffic systems (ITSs). Unfortunately, almost all current map-matching approaches were developed for GPS trajectories generated by probe sensors mounted in a few vehicles and cannot deal with the trajectories of massive vehicle samples recorded by fixed sensors, such as camera detectors. In this paper, we propose a novel map-matching model termed Fixed-MM, which is designed specifically for fixed sensor data. Based on two key observations from real-world data, Fixed-MM considers (1) the utility of each path and (2) the travel time constraint to match the trajectories of fixed sensor data to a specific path. Meanwhile, with the laws derived from the distribution of GPS trajectories, a path generation algorithm was developed to search for candidates. The proposed Fixed-MM was examined with field-test data. The experimental results show that Fixed-MM outperforms two types of classical map-matching algorithms regarding accuracy and efficiency when fixed sensor data are used. The proposed Fixed-MM can identify 68.38% of the links correctly, even when the spatial gap between the sensor pair is increased to five kilometers. The average computation time spent by Fixed-MM on one point is only 0.067 s, and we argue that the proposed method can be used online for many real-time ITS applications.

1. Introduction

Map matching is the process of correctly identifying the path on which a vehicle is travelling [1]. It provides a promising opportunity to upgrade the service level of various intelligent traffic system (ITS) applications [2–4]. However, the current map-matching algorithms are generally designed for satellite-based GPS points that are provided by probe sensors mounted on probe vehicles. These probe vehicles provide spatial traffic information and direct measurements of travel time to monitor the traffic conditions in a citywide road network.

However, probe sensor data have limitations. The cost of purchasing GPS units and transferring data can severely limit the scale of probe samples. Only a biased estimation of the traffic information can be obtained because the probe data are usually collected from one type of vehicle, such as taxis. Additionally, a probe sensor system imposes an enormous computational burden on the system administration owing to high polling frequency and positional noise [5].

Fixed sensor data show the potential to overcome the issues existing in the probe sensor data. Fixed sensors, such as cameras, loops, and microwaves, are widely used in urban traffic monitoring and management (with the development of ITS technology, camera sensors have been improved in terms of accuracy, cost, and ease of use. Therefore, the fixed sensor data considered in this paper refer specifically to the observations collected through camera-based sensors). The transit information of every vehicle approaching the fixed sensor station is captured. Consequently, the movement patterns of almost all vehicles running on a road network with fixed sensors can be recorded. This provides opportunities to reduce the estimation bias in traffic information. The fixed sensor system may also improve the efficiency of the map-matching process with a reduced polling frequency and more accurate location record, even for a large-scale urban traffic system.

Many map-matching methods have been developed, and their reviews can be found in [1, 6]. Quddus et al. classified the methods into four categories: geometric, topological, probabilistic, and advanced. However, such approaches can only perform well with high-frequency GPS data and may become less effective for low-frequency trajectory data [6]. In recent years, two groups of methods, namely, HMM-based algorithms and ST-Matching algorithms, have been developed to deal with the sparsity issue of low-polling frequency trajectory data.

(i) HMM-based algorithms: Newson and Krumm [7] introduced a two-step map-matching algorithm based on a hidden Markov chain for a sparse GPS trajectory, called the HMM algorithm. First, this method finds a set of candidate links for each GPS point and defines a measurement probability to describe how the GPS point is aligned with each candidate link. Then, it connects each pair of consecutive candidate links with the shortest path to generate the candidate graph. Next, a transition probability defines the likelihood of the tracking vehicle moving along each candidate path. Finally, the best matching path sequence is identified using the Viterbi algorithm. The experimental results show that even with sampling intervals of 30 s, the accuracy of this algorithm is barely degraded. However, it has high computational complexity and becomes slow when working with long trajectories and extended search radii. Mohamed et al. [8] employed three filters (i.e., speed, direction, and α-trimmed mean filters) to reduce the candidate sets for improving the efficiency of the map-matching process. Koller et al. [9] proposed a fast-HMM algorithm that replaces the Viterbi algorithm with the bidirectional Dijkstra to determine the optimal map-matching solution. This algorithm can avoid up to 45% of the costly routing operations without negatively affecting the map-matching result. Han et al. [10] partitioned road networks into approximate segments and then indexed the approximate segments into an optimised packed R tree to improve the road-network search duration. It has also been argued that mobility in a road network is non-Markovian. Jagadeesh and Srikanthan [11] complemented the HMM algorithms with the concept of drivers' route choice. The results show that this improves matching accuracy further, especially at high levels of noise.

(ii) ST-Matching algorithms: Lou et al. [12] introduced a map-matching algorithm for low-polling frequency GPS trajectories based on both spatial and temporal analysis, called ST-Matching. It modelled temporal analysis using speed and travel time data to improve its accuracy. The experimental results show that ST-Matching is more robust to the decrease in sampling rate than the map-matching algorithm using only spatial information, indicating that temporal constraints are indeed useful in map matching with sparse trajectory data. Considering that this method cannot handle the matching error well at junctions, Hsueh and Chen [13] introduced directional analysis to ST-Matching, called STD-Matching. It employs real-time directional motion with the directional analysis function to reflect the influence of a user's true movement over the GPS trajectories. The experimental results demonstrate that the STD-Matching algorithm significantly improves the matching accuracy. Liu et al. [14] proposed a spatial and temporal conditional random field map-matching method called the ST-CRF algorithm. The ST-CRF model considers both spatial and temporal accessibility between two GPS points, in addition to consistency in the direction of travel. A series of experiments showed that the ST-CRF method has better performance and robustness and solves the "label-bias" problem in the HMM algorithm.

The above-mentioned map-matching algorithms are mainly designed for low-frequency probe sensor data, such as GPS trajectories. They may become less effective for fixed sensor data because the fixed sensor data differ from probe sensor data in at least two aspects:

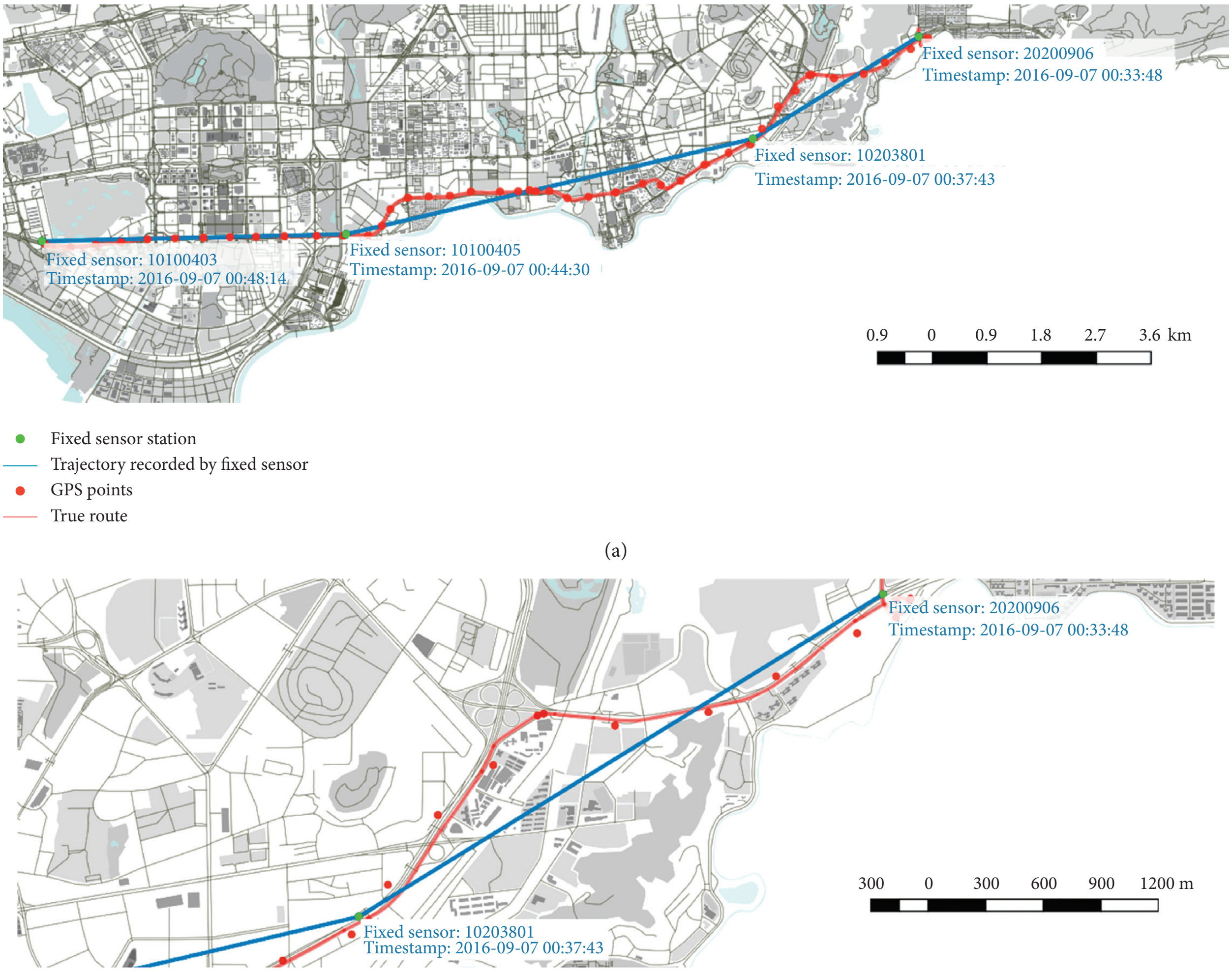
(a) The fixed sensor data are much sparser than the probe sensor data. As shown in Figure 1(a), the distance between consecutive points recorded by fixed sensors is usually dozens of times that recorded by the probe sensor. Hence, there are too many possible paths to be matched between neighbouring fixed sensors. If only the shortest path length is considered (as in the current map-matching algorithms developed for probe sensors), the realistic paths may not be adequately evaluated.

(b) The positions provided by the fixed sensors are fixed and accurate, while the probe sensors move along with the probe vehicle and generate GPS points with random errors [15]. Figure 1(b) presents a microscopic view of the trajectories between the fixed sensors 20200906 and 10203801. One easily finds that the fixed sensor data (green points) are located accurately on the road links, and the probe sensor data (red points) are always positioned several meters away from the true path.

In this study, we developed a map-matching algorithm designed specifically for fixed sensor data, called Fixed-MM. For this purpose, the conventional map-matching models for probe sensor data are abbreviated as Probe-MM. The contributions of Fixed-MM can be summarised as follows:

(a) It combines both route choice preferences and temporal constraints to identify the true path of the fixed sensor data. The experimental results show that the proposed method significantly improves the matching accuracy.

(b) Fixed-MM developed a candidate-path generation algorithm to search for a realistic path by relaxing the assumption that the location of each point is noisy. In this manner, the time-consuming candidate-path



(a)

(b)

Figure 1: Comparison of fixed sensor and probe sensor data. (a) Macroscopic and (b) microscopic views of one vehicle's trajectories recorded by the probe and fixed sensors.

generation process can be conducted separately and in parallel, and average computation time of the matching process for a point is reduced to 0.067 s.

The remainder of this paper is organised as follows. The problem definition and overview of the framework are presented in the Preliminaries section. Then, the Fixed-MM algorithm and candidate-path set generation algorithm are proposed in the Methodology section. The Experiment section details the process and presents the experimental results. Finally, we conclude the paper in the last section.

2. Preliminaries

2.1. Formulation of the Map-Matching Problem. To better illustrate Fixed-MM, the definitions of variables and the problem are introduced in this section.

Definition 1. Road network: a road network (RN) consists of a set of road links $\{l\}$ connected in a graph format. Each road link, l, is a directed edge with two terminal points, a length (l.len), a level (l.lev) (e.g., an expressway, a primary road, or a secondary road), a direction (l.di) (e.g., one-way or bidirectional), and the number of lanes (l.lan).

Definition 2. Path: path P is represented by a sequence of connected road links, P: l1, l2,. . ., lx,. . ., lX, in an RN.

Definition 3. Fixed sensor trajectory: a fixed sensor trajectory, Tr, is a sequence of time-ordered points, Tr: $F_{\mathrm{id}^{(1)}}$, $F_{\mathrm{id}^{(2)}}$, . . ., $F_{\mathrm{id}^{(j)}}$, . . ., $F_{\mathrm{id}^{(J)}}$, where each point $F_{\mathrm{id}^{(j)}}$ has a unique identification number, id, geospatial coordinate, $(F_{\mathrm{id}^{(j)}} \cdot \mathrm{lon}$, $F_{\mathrm{id}^{(j)}} \cdot \mathrm{lat})$, and timestamp, $F_{\mathrm{id}^{(j)}} \cdot t$.

Definition 4. Sensor pair: a sensor pair is two neighbouring points in a Tr, namely, $(F_{\mathrm{id}^{(j)}}, F_{\mathrm{id}^{(j+1)}})$, $j = 1, 2, \ldots, J–1$, where $F_{\mathrm{id}^{(j)}}$ is the original fixed sensor point and $F_{\mathrm{id}^{(j+1)}}$ is the destination fixed sensor point.

Definition 5. Candidate path set: the candidate path set, Φ_j, consists of all paths with a nonzero probability of matching between a given sensor pair $(F_{\mathrm{id}^{(j)}}, F_{\mathrm{id}^{(j+1)}})$, while all unrealistic paths have a probability of zero.

Now the problem of Fixed-MM is defined as follows.

Problem 1. Given a fixed sensor trajectory Tr and a road network RN, for each sensor pair $(F_{\mathrm{id}^{(j)}}, F_{\mathrm{id}^{(j+1)}})$ in Tr, find a path Pi from Φ_j with the highest probability of being a matched path.

2.2. Framework. The framework of Fixed-MM is illustrated in Figure 2. Three types of datasets, including fixed sensor data, probe sensor data, and road network data, are used as inputs. The trajectory of the fixed sensor data is first decomposed into separate sensor pairs. The probe sensor data are also matched with a specific path based on the Probe-MM algorithm. Meanwhile, a candidate path generation algorithm is used to search for possible paths for each sensor pair. Then, the matching probability for each candidate path is calculated, and the matching results can be attained by finding the candidate path with the highest matching probability.

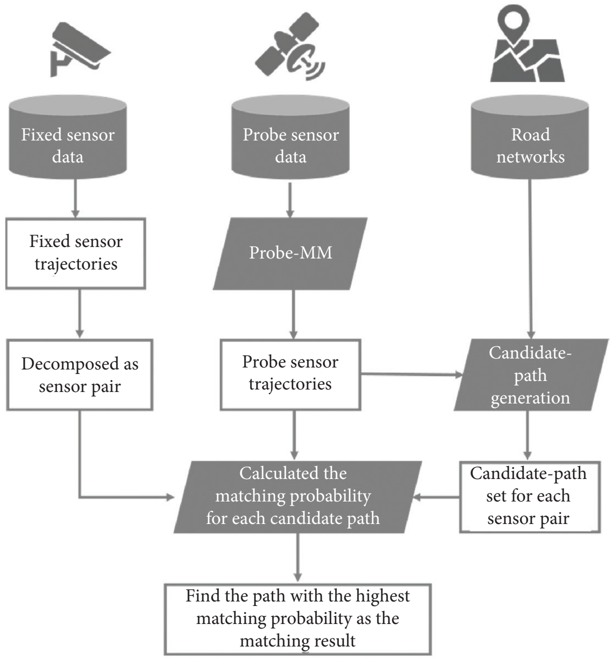

Fixed sensor data
Probe sensor data
Road networks
Fixed sensor trajectories
Probe-MM
Decomposed as sensor pair
Probe sensor trajectories
Candidate-path generation
Calculated the matching probability for each candidate path
Candidate-path set for each sensor pair
Find the path with the highest matching probability as the matching result

Figure 2: Framework of the proposed Fixed-MM Algorithm.

3. Methodology

3.1. Characteristics of the Data. The key to Fixed-MM is finding the most likely path to connect the sensor pair. In this section, we provide two key observations of the true trajectories that lead to the proposed approach. Figure 3(a) illustrates the GPS trajectory of 1365 sample vehicles travelling between the sensor pair $(F_{20507303}, F_{20501803})$, and they are taken as examples to illustrate the observed laws.

Observation 1. The drivers prefer to travel along the path with high utility.

Example 1. Consider path A, path B, and path C visualised in Figure 3(a) with their attributes summarised in Table 1. Sixty-eight percent of the samples travel path A, while only 32% of the samples travel along the other two. Thus, it is reasonable to infer that drivers prefer to choose paths with less travel time, fewer intersections, and more high-level road links, which indicates that the higher the utility of the path, the more attractive the path is to the driver.

Observation 2. The observed travel time tends to be close to the expected travel time of the true path.

Example 2. Based on the Prob-MM algorithm, the GPS trajectories can be matched to three paths. The histograms of the observed travel times for the three paths are calculated in Figure 3(b). It is easily found that the histograms fit well to the normal distribution, which means that a path's observed travel time tends to be close to its expected travel time (average travel time). If the observed travel time of a sample is 18 min, we may infer that this trip is very likely to be matched with path C.

Based on the above observations, we propose a novel map-matching algorithm for fixed sensor data, namely, Fixed-MM that incorporates both (1) the utility of each route and (2) the travel time constraint to identify the path with the highest probabilities from the candidate path set as the matched path. Details of the utility model, travel time constraint, and candidate path set generation algorithm are described in the following subsections.

3.2. Utility Model. Similar to the route choice model, the travel behaviour preference reflected in Observation 1 is modelled with utility theory. It assumes that the driver's preference for a path is captured by a value called utility, and the driver selects the path in the candidate set with the highest utility [16].

Let $U_{i,j}$ be the utility of the ith path $P_{i,j}$ belonging to the candidate set Φ_j of the sensor pair: $(F_{\mathrm{id}^{(j)}}, F_{\mathrm{id}^{(j+1)}})$. It consists of a deterministic term $V_{i,j}$ and a random term $\varepsilon_{i,j}$ such that

$$U_{i,j} = V_{i,j} + \varepsilon_{i,j}. \tag{1}$$

The random term $\varepsilon_{i,j}$ is assumed to be independent and identically distributed (i.i.d.) as a Gumbel distribution. The deterministic term is assumed to have a linear relationship with path attributes, such that

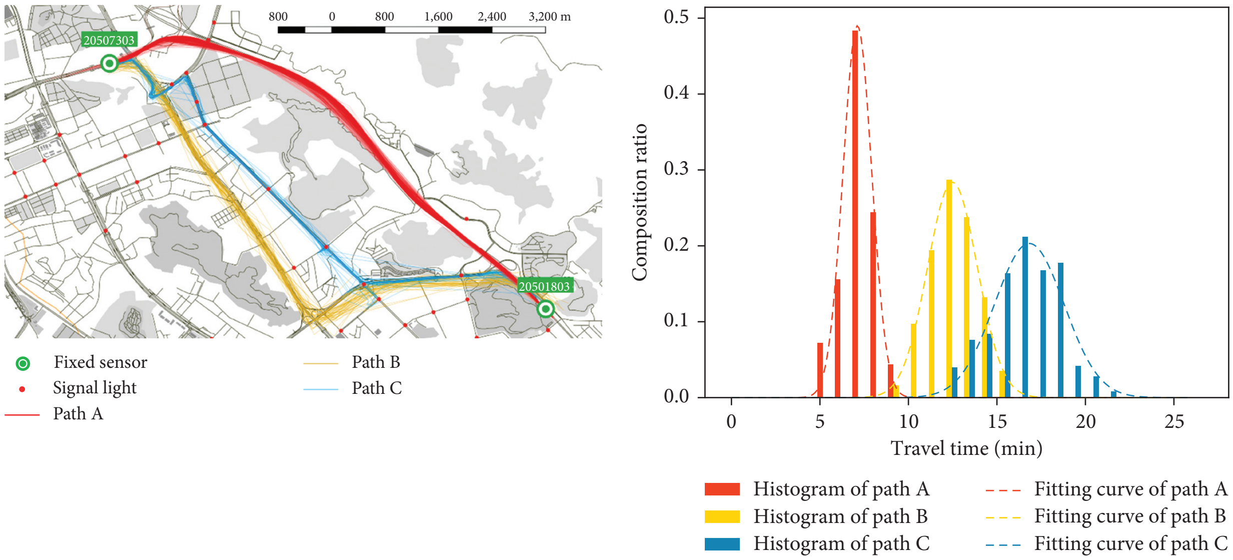
$$V_{i,j} = \beta^{\mathrm{FTT}} x_{i,j}^{\mathrm{FTT}} + \beta^{\mathrm{NSL}} x_{i,j}^{\mathrm{NSL}} + \beta^{\mathrm{PE}} x_{i,j}^{\mathrm{PE}}, \tag{2}$$

where $x_{i,j}^{\mathrm{FTT}}$, $x_{i,j}^{\mathrm{NSL}}$, and $x_{i,j}^{\mathrm{PE}}$ are vectors of the observed path attributes and β^{FTT}, β^{NSL}, and β^{PE} are vectors of coefficients that represent drivers' preferences on path attributes. The descriptions of the path attributes are presented in Table 2.

Based on the above definitions of path utility, the matching probability of a candidate path $P_{i,j}$ is given by [16]

$$\Pr\left(P_{i,j}\right) = \frac{e^{V_{i,j}}}{\sum_{P_{i',j}\in\Phi_j} e^{V_{i',j}}}. \tag{3}$$

Equation (3) can also be transformed as



(a)

(b)

Figure 3: The sensor pair example. (a) The location of the sensor pair example and the GPS trajectories. (b) The distribution of travel times.

Table 1: Attributes of path A and path B.

Attributes	Path A	Path B	Path C
Length of route (m)	8239.25	9022.70	8971.15
Number of signal lights	0	0	7
Average travel time (min)	7.11	12.49	16.83
Proportion of expressway	1	1	0.37

$$\Pr\left(P_{i,j}\right) = \frac{e^{V_{i,j}}}{\sum_{P_{i',j}\in\Phi_j} e^{V_{i',j}}} = \frac{1}{\sum_{P_{i',j}\in\Phi_j} e^{V_{i',j}-V_{i,j}}}. \tag{4}$$

It is easy to find that the larger the difference between the utility $V_{i,j}$ and the other $V_{i',j}$s, the higher the matching possibility, $\Pr(P_{i,j})$. This means that the candidate path with higher utility is more likely to be matched, which corresponds to the rule reflected in Observation 1.

3.3. Temporal Constraint. To consider Observation 2, the temporal constraint between the observed travel time and expected travel time of a candidate path must be modelled. Their definitions are as follows.

The observed travel time $t_{j,n}$ is the time spent by the nth sample when travelling between sensor pairs $(F_{\text{id}^{(j)}}, F_{\text{id}^{(j+1)}})$ and can be obtained by calculating the difference between the transit timestamps recorded by $F_{\text{id}^{(j)}}$ and $F_{\text{id}^{(j+1)}}$:

$$t_{j,n} = F_{\text{id}^{(j+1)}} \cdot t - F_{\text{id}^{(j)}} \cdot t. \tag{5}$$

The expected travel time $\widehat{t_{i,j}}$ is the average travel time of the candidate path, $P_{i,j}$, where $P_{i,j} \in \Phi_j$. This can be calculated based on probe sensor data:

$$\widehat{t_{i,j}} = \sum_{l_x\in P_{i,j}} \frac{\sum_{n=1}^{N} t_{x,n}}{N_x}, \tag{6}$$

where $t_{x,n}$ is the travel time spent by the n^{th} sample on road link l_x, and N_x is the total number of probe vehicles traversing road link l_x.

The temporal constraint can be calculated based on the deviation $t_{j,n} - \widehat{t_{i,j}}$ between the observed $t_{j,n}$ and the expected travel times, $\widehat{t_{i,j}}$. This is attributed to a combination of the natural variation in travel times and the error in the travel time estimate. The deviations of the three sample paths are shown in Figures 4–6 in Appendix A, respectively. The travel time varies significantly on different paths depending on the time of day, and all the histograms of $t_{j,n} - \widehat{t_{i,j}}$ during the morning peak fit well to the normal distribution. Therefore, we can assume that the deviations have a Gaussian distribution $t_{j,n} - \widehat{t_{i,j}} \sim N(\mu_s, \sigma_s)$. μ_s and σ_s are the mean and variance of $t_{j,n} - \widehat{t_{i,j}}$ for the candidate path $P_{i,j}$, during period s. Then, the temporal constraint $q(t_{j,n} - \widehat{t_{i,j}})$ can be defined as

$$q\left(t_{j,n} - \widehat{t_{i,j}}\right) = \frac{e^{-0.5\left(t_{j,n}-\widehat{t_{i,j}}-\mu_s/\sigma_s\right)^2}}{\sum_{P_{i',j}\in\Phi_j} e^{-0.5\left(t_{j,n}-\widehat{t_{i',j}}-\mu_s/\sigma_s\right)^2}}. \tag{7}$$

The denominator aims at normalizing the temporal constraint to one.

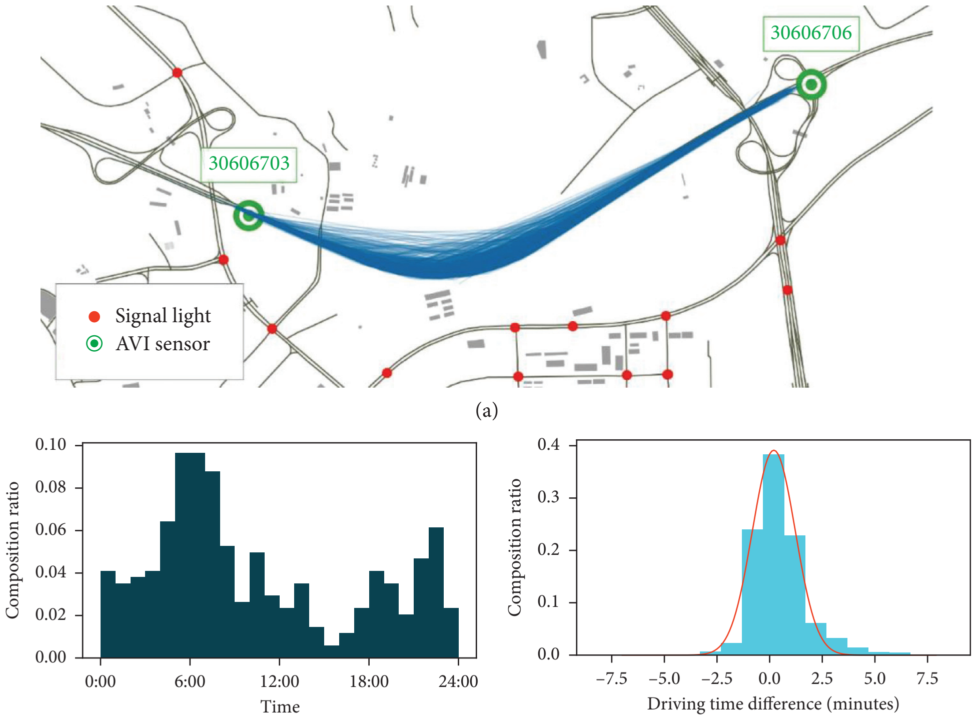
We added the temporal constraint as a correction term for the utility function. Then, the matching probability can be rewritten as

$$\Pr\left(P_{i,j}\right) = \frac{e^{V_{i,j}+\alpha\ln q\left(t_{j,n}-\widehat{t_{i,j}}\right)}}{\sum_{P_{i',j}\in\Phi_j} e^{V_{i',j}+\alpha\ln q\left(t_{j,n}-\widehat{t_{i',j}}\right)}}, \tag{8}$$

where α is a scale parameter. The correct term $\alpha \ln q(t_{j,n} - \widehat{t_{i,j}})$ in equation (8) describes the likelihood of compliance between the observed $t_{j,n}$ and expected travel time $\widehat{t_{i,j}}$. When $t_{j,n} - \widehat{t_{i,j}}$ is smaller (the observed travel time is closer to the expected travel time), $q(t_{j,n} - \widehat{t_{i,j}})$ becomes larger.

Table 2: Attributes used for path utility.

Attributes	Parameters	Descriptions
$x_{i,j}^{\mathrm{NSL}}$	β^{NSL}	Number of signal lights (NSL): the number of signal lights along the path
$x_{i,j}^{\mathrm{FTT}}$	β^{FTT}	Free travel time (FTT): free flow travel time along the path (unit: s)
$x_{i,j}^{\mathrm{PE}}$	β^{PE}	The proportion of expressway (PE): the proportion of expressway links



(a)

(b)

(c)

Figure 4: (a) GPS trajectories of samples. (b) Temporal distribution of the samples. (c) Histogram of the temporal constraint between 6:00 and 7:00 (fitting result: $\mu = 0.20$ and $\sigma = 1.02$, goodness of fit: 0.99).

According to equation (4), the matching probability increases $P_{i,j}$. This is also in line with Observation 2 in the previous section.

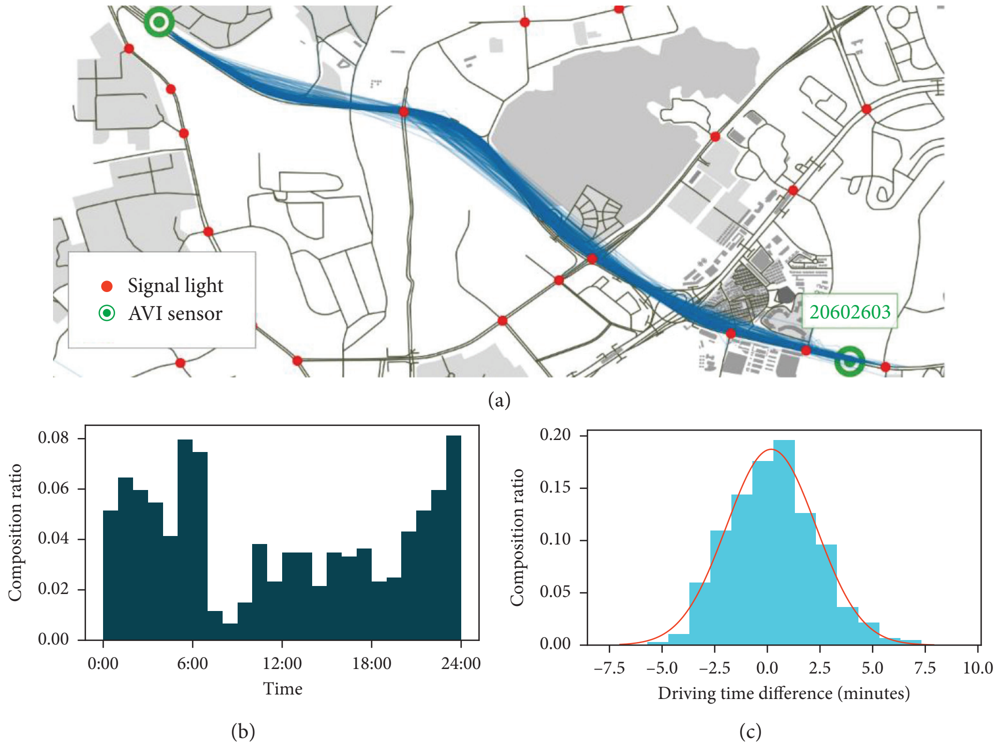
3.4. Generating Candidate Path Set. Finding all possible paths that connect each sensor pair as candidates is another key step for Fixed-MM. The candidate path set is usually large, with a long distance between the paired sensors, and a dense urban road network. In addition, preferential and realistic paths should be included because comparing a path to a set of highly unattractive and unrealistic candidates would not provide much useful information [17]. In this study, we develop a protocol for generating a realistic candidate path set based on the following observations:

Observation 3. There may be certain detours on the candidate paths.

Example 3. Figure 7(a) illustrates the GPS trajectories of 620 samples that travel between sensors $F_{20507301}$ and $F_{20507302}$ near the Bao'an International Airport in Shenzhen, China. Based on the map-matching algorithm designed for the probe data, each GPS point was projected onto a specific link. The observed number of samples on each link is represented by different colours in Figure 7(b). Most (92%) of the samples have a large offset against the shortest path, and the departure platform of the airport was chosen as a destination on the way. This indicates that there may be certain detours on these popular paths. These circuitous paths may be considered as unattractive alternatives for route choice models. However, they are popular candidates in the context of map-matching algorithms.

Observation 4. Trajectories captured by a sensor pair will not pass the links monitored by other fixed sensors.

Example 4. As shown in Figure 7(a), the road link monitored by the fixed sensor $F_{20507403}$ has never been travelled by any vehicle captured by the sensor pair $(F_{20507301}, F_{20507302})$. The reason for this phenomenon is that if a vehicle has travelled on the link where $F_{20507403}$ located, the pass information will be recorded, and then the sensor pair

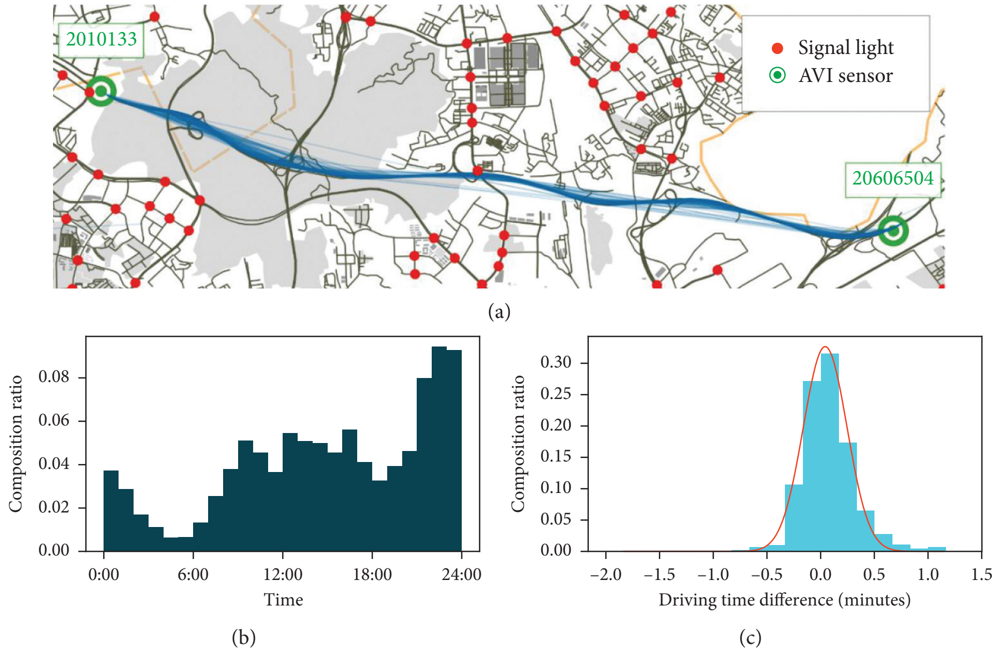


(a)

(b)

(c)

Figure 5: (a) GPS trajectories of samples. (b) Temporal distribution of the samples. (c) Histogram of the temporal constraint between 6:00 and 7:00 (fitting result: $\mu = 0.18$ and $\sigma = 2.13$, goodness of fit: 0.98).



(a)

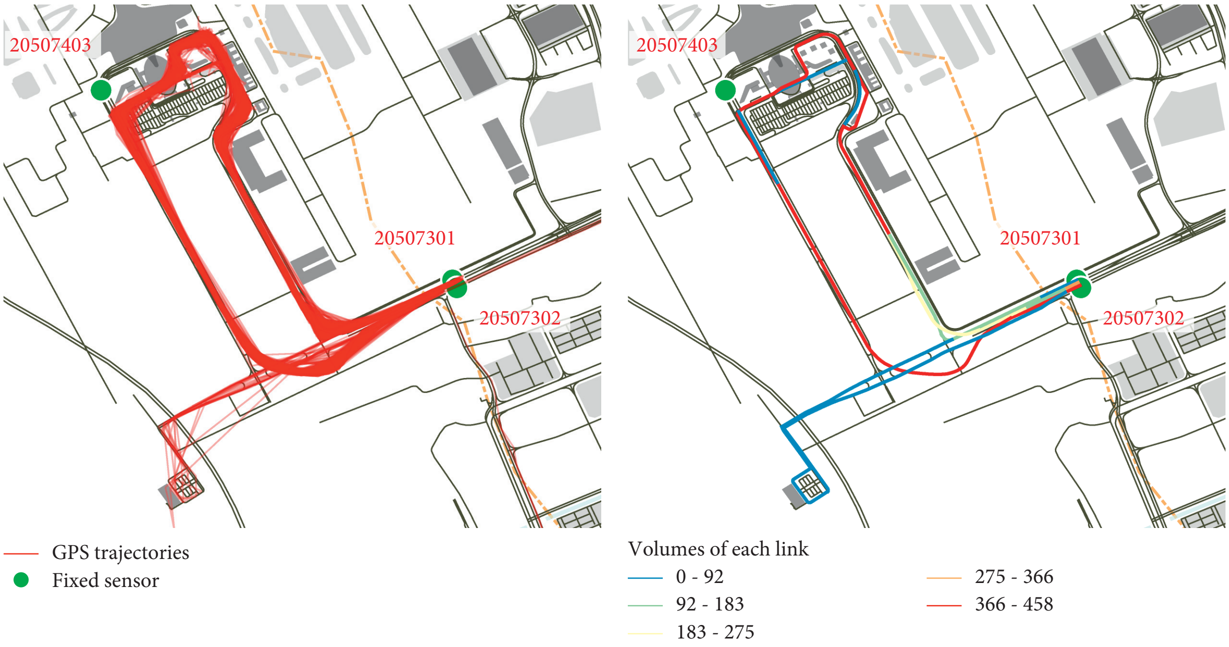
(b)

(c)

Figure 6: (a) Trajectories of samples. (b) Temporal distribution of the samples. (c) Histogram of the temporal constraint between 6:00 and 7: 00 (fitting result: $\mu = 0.25$ and $\sigma = 1.22$, goodness of fit: 0.99).

$(F_{20507301}, F_{20507302})$ will be decomposed into two sensor pairs, namely, $(F_{20507301}, F_{20507403})$ and $(F_{20507403}, F_{20507302})$.

In this paper, we believe that historical GPS trajectories contain useful information about the composition of popular candidates. Thus, the candidate path does not necessarily conform to behavioural assumptions but must be realistic; we use a biased random walk algorithm, which was first proposed by [17] to generate the candidate set. It draws a candidate path



(a)

(b)

Figure 7: Example of realistic path generation. (a) GPS trajectories. (b) Volumes of each link.

through a succession of random turns. The pseudocode of the candidate set generation algorithm is presented in Algorithm 1. The key to this algorithm is how the probability of turning is defined. In contrast to the original random walk algorithm, we set the turning probability of the links where other fixed sensors are located at 0 to satisfy the rule contained in Observation 4. In other situations, the turning probability is calculated based on field-test probe sensor data rather than the shortest path assumption. In this manner, the candidate path with the destination described in Observation 3 can be generated.

Based on the above analysis, the turning probability is defined as

$$\Pr\left(l_x, l_y\right) = \begin{cases} 0, & l_y \in \Phi_{\mathrm{FS}} \text{ and } l_y \neq l_s, l_e, \\ \dfrac{N_{xy}}{\sum_{l_{y'} \in \Phi_x} N_{xy'}}, & \text{otherwise,} \end{cases} \tag{9}$$

where Φ_{FS} is the set of links monitored by the fixed sensors, l_s is the start link where the origin fixed sensor is located, l_e is the end link where the destination fixed sensor is located, N_{xy} is the number of GPS trajectories traversing from link l_x to l_y, and Φ_x is the set of outgoing links that connect the sink link l_x.

4. Experiment

4.1. Experimental Dataset. To examine the proposed Fixed-MM algorithm, both fixed and probe sensor data were used with the basic digital road network.

Road Network: the shapefile of the road network in Shenzhen, China, was used [18]. The network graph contained 237,440 vertices and 215,771 road links. As shown in Figure 8, the road network covers a 40 × 50 km spatial area, with a total length of 21,985 km.

Fixed sensor dataset: A fixed sensor dataset generated by 715 cameras in Shenzhen from September 1 to October 31, 2016, was used. The transit information of vehicles was recorded, including license plate, timestamp, and detector ID.

Probe sensor dataset: we used a GPS trajectory dataset generated by 14,230 taxicabs during the same time range (from September 1 to October 31, 2016) as a probe sensor dataset. The GPS records include license plates, timestamps, and coordinates. The average sampling rate was set at 15 seconds per point.

With identical license plate information, we can extract the probe and fixed sensor data of the same taxicab as observed samples to train and test our model.

In the implementation, we removed noncontinuous driving trips. The main reason is that this noncontinuous driving part of the sample trips contains great uncertainty and will increase the estimation error of the Fixed-MM. Finally, 1,485,476 samples were extracted as a training dataset for estimation, while 156,192 samples were used as the testing dataset for evaluation. The estimation and evaluation of the Fixed-MM are introduced in the following sections.

4.2. Model Estimation. The coefficients of the Fixed-MM reflect the matching results' sensitivity to the variables. The

```
Input: The road network $RN$ and the link pair $(l_s, l_e)$, where $l_s$ and $l_e$ are the links where the origin fixed sensor $F_{\mathrm{id}^{(j)}}$ and destination fixed sensor $F_{\mathrm{id}^{(j+1)}}$ are located.
Output: The candidate set $\Phi_j$ for sensor pair $(F_{\mathrm{id}^{(j)}}, F_{\mathrm{id}^{(j+1)}})$.
Initialization
  Set the candidate set: $\Phi_j = \varnothing$
  Set the size of the candidate set: $DN$
Turning Probability
  For $l_x$ in road network $RN$:
    Calculate the turning probability $\Pr(l_x, l_y)$ based on equation (9).
Random Walk
  While $n < DN$ do
    $l_x = l_s$
    $P = [l_s]$
    While $l_y \neq l_e$ do
      Randomly select a next link $l_y$ based on the turning probability $\Pr(l_x, l_y)$
      Update the generated path: $P$.append$(l_y)$
      Update the current link: $l_x = l_y$
    End while
    $n += 1$
    Update the candidate set: $\Phi_j = \Phi_j \cap P$
  End while
```

Algorithm 1: Candidate set generation algorithm.

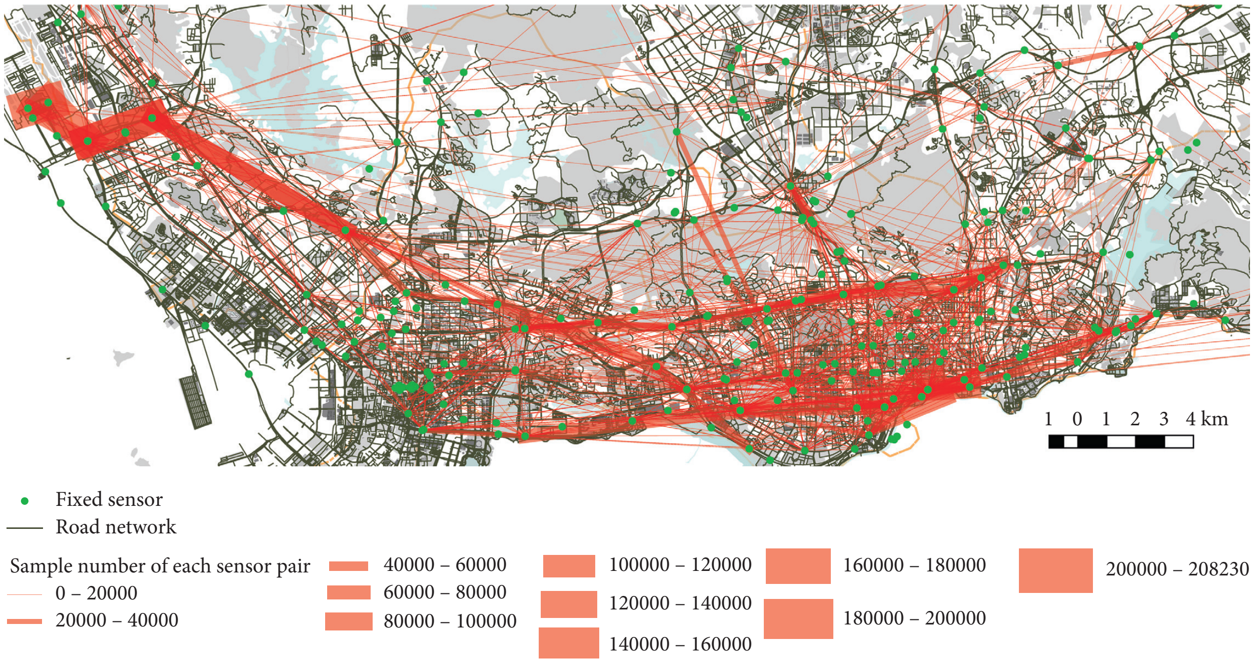


Figure 8: Distribution of observed trips.

values of the unknown parameters based on the training dataset must be identified. In this study, we consider the most widely used estimation procedure: the maximum likelihood technique [19].

Given the high number of sensor pairs, it is impossible to present detailed estimation results for each pair. Therefore, we only provide the detailed estimation results of the example sensor pair: $(F_{2010002}, F_{1010403})$. The GPS trajectories of the samples between this sensor pair are shown in Figure 9 in Appendix B. The candidate path set generated by the algorithm proposed in this paper is illustrated in Figure 10.

Both the Fixed-MM model without temporal constraints (defined by equation (3)) and the Fixed-MM with temporal constraints (defined by equation (8)) are estimated. The estimation results of the two models are presented in Table 3, and several findings can be obtained.

Finding 1: as expected, the estimated parameter of "free travel time" and "number of signal lights" has a

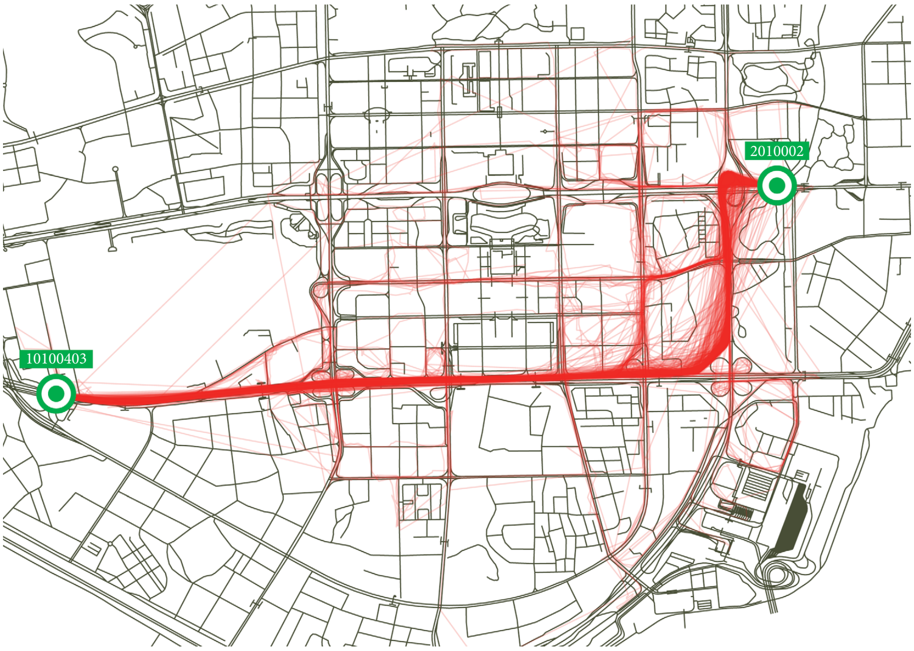


Figure 9: GPS trajectories between the example sensor pair: ($F_{2010002}$, $F_{10100403}$).

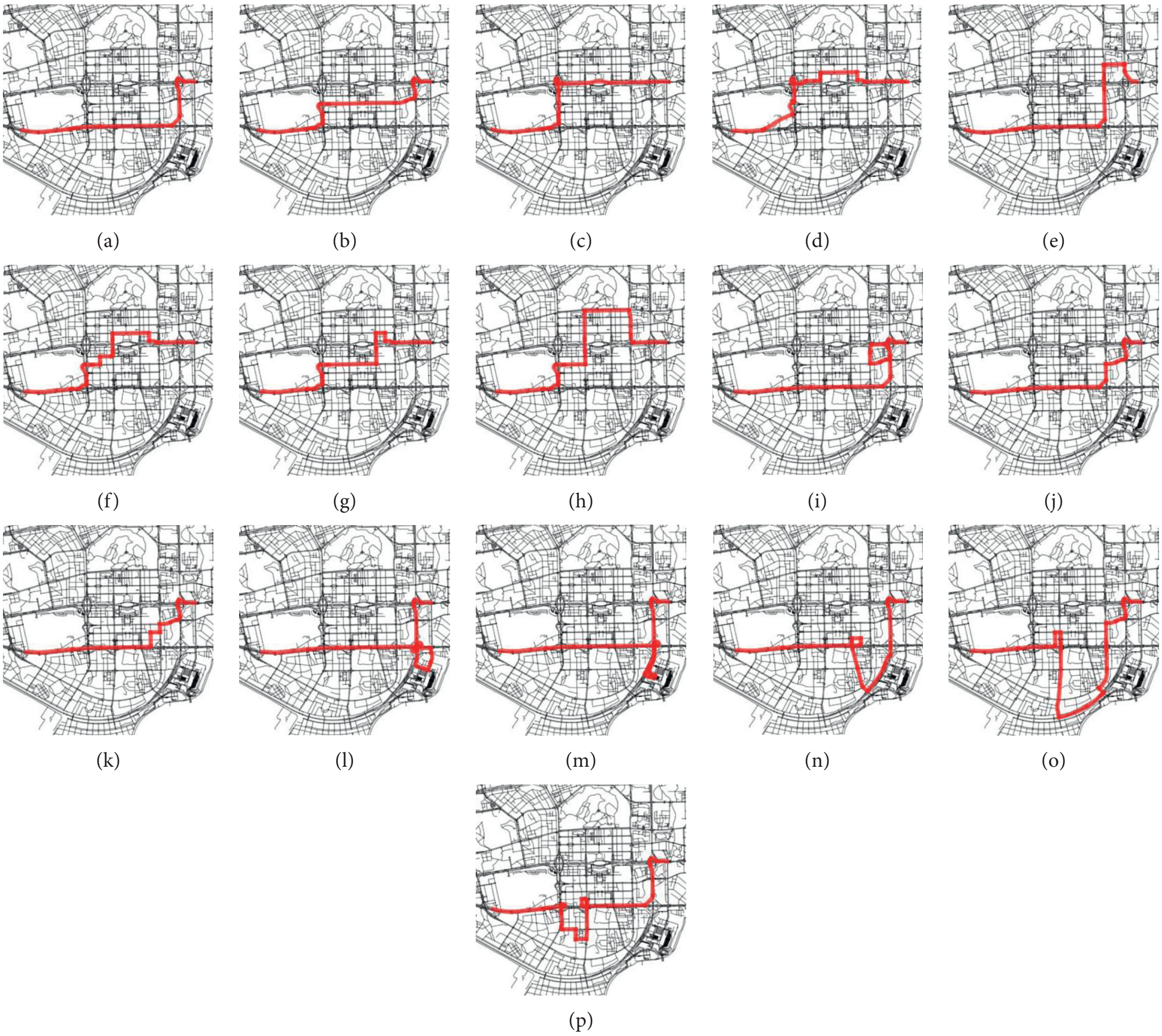


Figure 10: (a–p) Generated candidate paths between the example sensor pair: ($F_{2010002}$, $F_{10100403}$).

Table 3: Estimation results of Fixed-MM.

Parameters	Fixed-MM (without temporal constraint)	Fixed-MM (with temporal constraint)
Number of signal lights β^{NSL}	−0.185	−0.188
t-test	−21.442	−11.579
Free travel time β^{FTT}	−0.273	−0.101
t-test	−9.574	−1.839
Proportion of expressway β^{PE}	0.074	4.140
t-test	49.532	10.925
Temporal constraint α	—	67.395
t-test	—	12.417
Log-likelihood	−17851.898	−1049.901
Adjusted likelihood ratio	0.511	0.971

negative sign and the "proportion of expressway" has a positive sign in each case. The negative sign and t-statistic of β^{NSL} and β^{FTT} suggest that the freer travel time and signal lights the path has, the less likely it is to be matched. The positive sign and t-statistic of β^{PE} imply that a path with a higher proportion of expressways will be more attractive to travellers.

Finding 2: the temporal constraint parameter, α, is very large, which means that the correct term has a significant effect on the matching results.

Finding 3: when the temporal constraint term, $\alpha \ln q(t_{j,n} - \widehat{t_{i,j}})$, was considered, the Fixed-MM model with temporal constraints had a much lower log-likelihood. Thus, we can infer that it has a better model fit and is closer to the true model.

4.3. Model Evaluation. In this section, we describe our algorithm on the testing dataset. Two classical Probe-MM algorithms are used as benchmarks, details of which are introduced as follows:

HMM algorithm [7]: given that the positions of the fixed sensors are located without noise, the measurement probability is set to 1 and only the transmission probability is considered

ST-Matching algorithm [12]: similar to the HMM-based algorithm, the observation probability in the spatial analysis of this method was set to 1 because of the accurate positions of the fixed sensors

In this study, two indexes for expressing matching accuracy were used. One is the accuracy length ratio of paths (ALRP) index, defined as follows:

$$\mathrm{ALRP} = \frac{\sum_{l_x \in P_{i,j}} \delta_x l_x \cdot \mathrm{len}}{P_{\mathrm{true}} \cdot \mathrm{len}} \times 100, \qquad (10)$$

where $l_x \cdot \mathrm{len}$ is the length of link l_x in the matched path, $P_{i,j}$ $P_{\mathrm{true}} \cdot \mathrm{len}$ is the total length of the true path, and $\delta_x = 1$ if l_x is also in the true path, and otherwise is 0.

The other index is the accuracy number ratio of paths (ANRP) index, which is defined as

$$\mathrm{ANRP} = \frac{\sum_{n=1}^{N_x} \delta_x}{N_x} \times 100, \qquad (11)$$

where N_x is the total number of links in the true path P_{true}.

Figures 11(a) and 11(b) show the ALRR and ANRR of the proposed Fixed-MM algorithm and two classical Probe-MM algorithms with regard to the spatial gap between fixed sensors. It can be seen clearly that our Fixed-MM outperforms both HMM and ST-Matching significantly. Meanwhile, the performance of two Probe-MM algorithms degrades sharply when the spatial gap decreases while Fixed-MM is more robust to the change of spatial gap. The proposed Fixed-MM can correctly identify 68.38% of the links, even when the spatial gap between the sensor pair increases to 5 km.

Because the candidate generation process and model training process can be conducted separately and in parallel, a comparison of the latency of the matching process may be more meaningful for online applications. In this study, the computation time for one point (ACTOP) was used to measure the computational latency of the map-matching algorithm.

As shown in Figure 12, the ACTOP of the two Probe-MM approaches increases dramatically as the spatial gap between the fixed sensors increases. Conversely, the ACTOP of Fixed-MM increases slowly. The main reason, therefore, can be deduced from two factors. The HMM and ST-MM algorithms assume that the position of the sensor is stochastic and noisy, and the candidate set must be regenerated for every sensor pair. It involves several shortest path computations between states at the previous and current time steps, which consumes most of the computation time. Conversely, the candidate set generation of the proposed method can be run in parallel and does not increase the computation time because the projection of the fixed sensor data is known and fixed. In

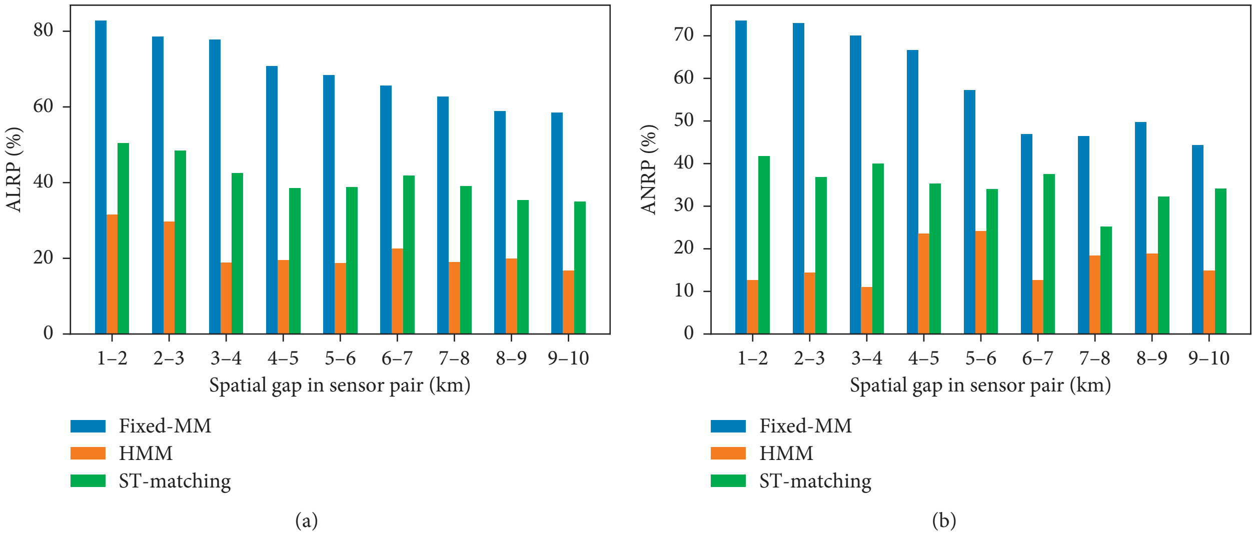


Figure 11: (a) ALRP and (b) ANRP with respect to the spatial gap.

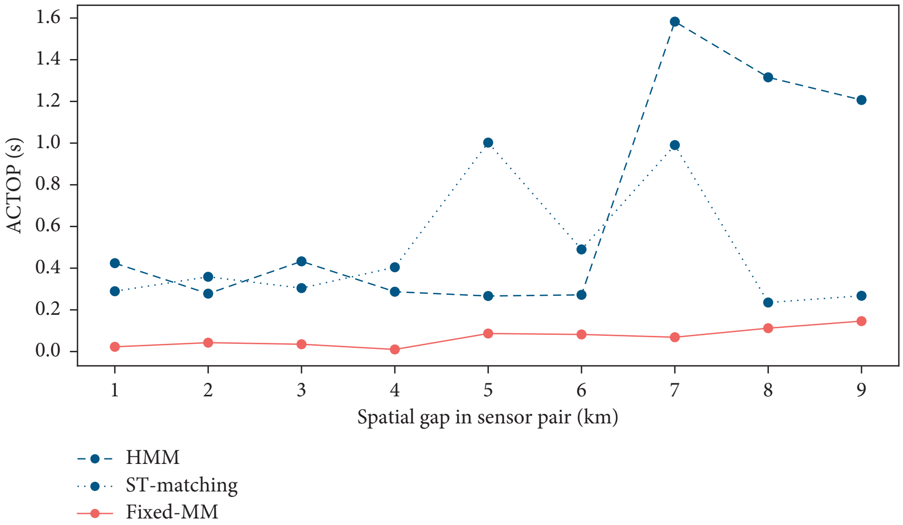


Figure 12: ACTOP with respect to the spatial gap (all algorithms are run on a Mac with 2.2 GHz Intel CPU and 16 GB RAM).

fact, the average ACTOP of Fixed-MM is only 0.067 s, and we argue that Fixed-MM can be performed online for many real-time ITS applications.

5. Conclusions

In this paper, we proposed a new map-matching algorithm called Fixed-MM to match vehicle trajectories recorded by fixed sensors onto a digital map. First, utility theory was employed to model the traveller's behaviour preference. Second, Fixed-MM was modified by adding a travel-time constraint term based on the observed and expected travel times. Moreover, a candidate path generation algorithm was designed for Fixed-MM.

Fixed sensor data and probe sensor data were collected as the experimental dataset. Both the Fixed-MM without a temporal constraint and Fixed-MM with a temporal constraint were estimated. The statistical results of the estimated parameters prove that the path attributes correlate significantly to the true path, and the Fixed-MM with the temporal constraint having a better model fit. The Fixed-MM algorithm was also compared with two classical Probe-MM algorithms in terms of matching accuracy and computational efficiency. Fixed-MM outperforms the two Probe-MM algorithms in both number (ANRR) and length (ALRR) accuracy indexes. Meanwhile, the Fixed-MM is more robust to changes in the spatial gap between fixed sensors. Fixed-MM also has a huge improvement in computing efficiency and exhibits potential for online applications. The experimental results demonstrate that the proposed Fixed-MM algorithm is both effective and efficient.

More research is needed in the future to determine the potential application value of Fixed-MM. Although the travel time and speed can also be estimated by the Probe-

MM algorithm with probe sensor data, the Fixed-MM provides a more diverse and credible estimation of travel time and speed. This is because the fixed sensor data covers almost all types of vehicles using the road network, while the probe sensor data can only be collected from one type of vehicle, for example, taxicabs. Meanwhile, with the application of Fixed-MM, more traffic information can be mined from the fixed sensor data. If all the observed trips of every fixed sensor can be matched to the road network, the traffic volumes of each path or link can be estimated, which is the key input value for traffic planning and management. Thus, our next research focus is to utilise the Fixed-MM to mine more reliable and accurate traffic state information from fixed sensor data. Moreover, since the fare gate in the AFC system is fixed, applying the proposed map-matching algorithm to learn the route choice behavior of subway passengers [20, 21] also presents great practical application values and is worthy of further study.

Appendix

A. Estimated Results of Temporal Constraint

GPS trajectories of samples are presented in Figures 4, 5, and 6.

B. Generated Candidate Path Set

GPS trajectories between the example sensor pair and generated candidate paths between the example sensor pair are presented in Figures 9 and 10, respectively.

Data Availability

The data used to support the findings of this study are available from the author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The authors are grateful to the Shenzhen Urban Transport Center for the research data and the Big Data Center of Southeast University for facilitating the numerical calculations in this paper. This work was supported in part by the National Key Research and Development Program of China under Grant 2019YFB1600200, the National Natural Science Foundation of China under Grant 71971056, 16th Regular Meeting Communication Program of China-Bulgaria Science and Technology Cooperation Committee under Grant no. 16-4, the Science and Technology Project of Jiangsu Province, China, under Grant BZ2020016, and Jiangsu Province Graduate Research and Practice Innovation Plan under Grant 101010573.

References

[1] M. A. Quddus, W. Y. Ochieng, and R. B. Noland, "Current map-matching algorithms for transport applications: state-of-the art and future research directions," *Transportation Research Part C: Emerging Technologies*, vol. 15, no. 5, pp. 312–328, 2007.

[2] K. Satyanarayana, A. D. Sarma, J. Sravan, M. Malini, and G. Venkateswarlu, "GPS and GPRS based telemonitoring system for emergency patient transportation," *Journal of Medical Engineering*, vol. 2013, Article ID 363508, 2013.

[3] S. An, L. Wang, H. Yang, and J. Wang, "Discovering public transit riders' travel pattern from GPS data: a case study in harbin," *Journal of Sensors*, vol. 2017, Article ID 5290795, 2017.

[4] M. N. Borhan, D. Syamsunur, N. Mohd Akhir, M. R. Mat Yazid, A. Ismail, and R. A. Rahmat, "Predicting the use of public transportation: a case study from putrajaya, Malaysia," *The Scientific World Journal*, vol. 2014, Article ID 784145, , 2014.

[5] T. Miwa, D. Kiuchi, T. Yamamoto, and T. Morikawa, "Development of map matching algorithm for low frequency probe data," *Transportation Research Part C: Emerging Technologies*, vol. 22, pp. 132–145, 2012.

[6] M. Hashemi and H. A. Karimi, "A critical review of real-time map-matching algorithms: current issues and future directions," *Computers, Environment and Urban Systems*, vol. 48, pp. 153–165, 2014.

[7] P. Newson and J. Krumm, "Hidden Markov map matching through noise and sparseness," in *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, Seattle, WA, USA, November 2009.

[8] R. Mohamed, H. Aly, and M. Youssef, "Accurate and efficient map matching for challenging environments," in *Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, IEEE, Dallas, TX, USA, November 2014.

[9] H. Koller, P. Widhalm, M. Dragaschnig, and A. Graser, "Fast hidden Markov model map-matching for sparse and noisy trajectories," in *Proceedings of the 2015 IEEE 18th International Conference on Intelligent Transportation Systems*, pp. 2557–2561, IEEE, Gran Canaria, Spain, September 2015.

[10] J. Han, X. Fu, L. Liu, and D. Jiang, "Online map matching by indexing approximate road segments," in *Proceedings: 2011 IEEE 2nd International Conference on Software Engineering and Service Science ICSESS 2011*, IEEE, Beijing, China, July 2011.

[11] G. R. Jagadeesh and T. Srikanthan, "Online map-matching of noisy and sparse location data with hidden Markov and route choice models," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 9, pp. 2423–2434, 2017.

[12] Y. Lou, C. Zhang, Y. Zheng, X. Xie, W. Wang, and Y. Huang, "Map-matching for low-sampling-rate GPS trajectories," in *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, Seattle, WA, USA, November 2009.

[13] Y.-L. Hsueh and H.-C. Chen, "Map matching for low-sampling-rate GPS trajectories by exploring real-time moving directions," *Information Sciences*, vol. 434, pp. 55–69, 2018.

[14] X. Liu, K. Liu, M. Li, and F. Lu, "A ST-CRF map-matching method for low-frequency floating car data," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 5, pp. 1241–1254, 2017.

[15] S. M. Saab and Z. M. Kassas, "Power matching approach for GPS coverage extension," *IEEE Transactions on Intelligent Transportation Systems*, vol. 7, no. 2, pp. 156–166, 2006.

[16] M. Ben-Akiva and M. Bierlaire, "Discrete choice methods and their applications to short term travel decisions," *Handbook of Transportation Science*, vol. 26, 2000.

[17] E. Frejinger, M. Bierlaire, and M. Ben-Akiva, "Sampling of alternatives for route choice modeling," *Transportation Research Part B: Methodological*, vol. 43, no. 10, pp. 984–994, 2009.

[18] S. Coast, "Open street map data extracts, open street map community," 2019, http://download.geofabrik.de/.

[19] K. E. Train, *Discrete Choice Methods with Simulation*, Cambridge University Press, Cambridge, UK, 2009.

[20] C. Yu, H. Li, X. Xu, and J. Liu, "Data-driven approach for solving the route choice problem with traveling backward behavior in congested metro systems," *Transportation Research Part E: Logistics and Transportation Review*, vol. 142, pp. 1–27, 2020.

[21] X. Xu, L. Xie, H. Li, and L. Qin, "Learning the route choice behavior of subway passengers from AFC data," *Expert Systems with Applications*, vol. 95, pp. 324–332, 2018.

Hindawi
Journal of Advanced Transportation
Volume 2021, Article ID 6646768, 14 pages
https://doi.org/10.1155/2021/6646768

Research Article

Analysis of Travel Hot Spots of Taxi Passengers Based on Community Detection

Shuoben Bi,[1] Yuyu Sheng,[1] Wenwu He,[2] Jingjin Fan,[3] and Ruizhuang Xu[1]

[1] *School of Geographical Sciences, Nanjing University of Information Science & Technology, No. 219 Ningliu Road, Nanjing 210044, China*
[2] *Mathematics and Physics Institute, Fujian University of Technology, No. 33 Xuefu South Road, University New District, Fuzhou 350001, China*
[3] *Institute for the History of Science and Technology, Nanjing University of Information Science & Technology, No. 219 Ningliu Road, Nanjing 210044, China*

Correspondence should be addressed to Shuoben Bi; bishuoben@163.com

Received 12 November 2020; Revised 6 March 2021; Accepted 16 March 2021; Published 26 March 2021

Academic Editor: Xinyue Xu

It is an important content of smart city research to study the activity track of urban residents, dig out the hot spot areas and spatial interaction patterns of different residents' activities, and clearly understand the travel rules of urban residents' activities. This study used community detection to analyze taxi passengers' travel hot spots based on taxi pick-up and drop-off data, combined with multisource information such as land use, in the main urban area of Nanjing. The study revealed that, for the purpose of travel, the modularity and anisotropy rate of the community where the passengers were picked up and dropped off were positively correlated during the morning and evening peak hours and negatively correlated at other times. Depending on the community structure, pick-up and drop-off points reached significant aggregation within the community, and interactions among the communities were also revealed. Based on the type of land use, as passengers' travel activity increased, travel hot spots formed clusters in urban spaces. After comparative verification, the results of this study were found to be accurate and reliable and can provide a reference for urban planning and traffic management.

1. Introduction

With the rapid development of information technology, spatial analysis driven by data forces geographic information science to face new challenges. Furthermore, visual analysis combined with geographic computing has greatly improved people's ability to mine new knowledge [1]. On the one hand, mobile information collection technology based on the global positioning system has become more mature; on the other hand, the flow space with urban residents' activities as the main carrier has become more extensive [2]. Although the intension of geographic information science has not changed, its content and form have become richer. Therefore, breaking through the traditional urban spatial research model is the key to discovering the law of urban residents' activities.

Time-space analysis based on residents' activities can explain the homogeneity of the influence of individual residents' behaviors on urban space, and the behaviors between different individuals can reflect that they are restricted by urban space and show their differences [3]. Therefore, as Harvey and Han [4] proposed the concept of geographic data mining and knowledge discovery, scholars have continued to explore knowledge in recent decades, and geography has experienced transition from an empirical paradigm to a system simulation paradigm and then to a data-driven paradigm [5]. Early research on the behavioral patterns of urban residents' activities mainly focused on extracting residents' activity points and on the correlation analysis of those points. For example, Veloso et al. [6] studied the strong association pattern of residents' activity locations, and Ahas et al. [7] studied the time difference and spatial distribution of residents' activities. Recent research has mainly focused on the identification of urban hot spot functional areas, urban accessibility analysis, urban boundary division, polycentric evaluation, etc. For example, Scholz et al. [8] studied urban residents' behavioral patterns and the temporal and spatial development of urban hot spots, and Cui et al. [9] studied the accessibility of urban residential areas and the distribution of low-access residential areas. Zhong et al. [10] studied the overall spatial structure of changes in the center of the city boundary, and Huang et al. [11] studied the effects of urban traffic and pedestrian activities. For two-layer fine-grained networks, Guo et al. [12] studied the structure of different urban road networks and developed corresponding datasets. Hamedmoghadam et al. [13] studied the displacement index of individual travel granularity to simplify collective behavioral patterns. In addition, there are related studies on urban planning and environmental safety assessment. For example, Zheng et al. [14] studied the characteristics of cross connectivity between urban planning and taxi driving, and Wu et al. [15] studied the temporal and spatial patterns of urban road traffic accidents.

In summary, the early research model was relatively narrow in its scope, only considering residents' activities but ignoring the characteristics of urban space. In recent years, research has become relatively rich, mainly based on urban planning, which was based on the analysis of residents' historical activities, such as behavioral patterns. The landmark research achievements are the GN algorithm [16] and the Newman fast algorithm [17], both of which are classic community detection algorithms. These can fully reveal the different resident activities, the spatial pattern, and the impact of potential factors on decision-making. In addition, Qin et al. [18] studied the traffic intensity and edge weight of network nodes based on the network interaction characteristics of urban hot spots.

The movement trajectory as a type of multisource sensor data has been widely adopted by researchers. Through the movement trajectory, the travel mode of residents' activities can be understood more clearly, the hot spots of activities can be extracted more accurately, and the reasons for the resident movement can be analyzed. Research on moving trajectories in Nanjing mainly includes Xu et al. [19], who found that traffic hot spots in Nanjing have the spatial distribution characteristics of agglomeration from the surroundings to the center; Yang et al. [20], who found that the Nanjing public transport system has the characteristics of cascading failure congestion; and Jin and Xu [21], who showed that the traffic flow on the key nodes of different grades of road network in Nanjing has obvious hierarchical structure characteristics. Therefore, this study aims to use the passenger pick-up and drop-off points extracted from taxi movement trajectories to explore the travel rules of taxi passengers, analyze the time and space patterns of taxi pick-up and drop-off communities, establish passenger travel activity indicators based on community detection [16], combine the data of graded roads and points of interest, explore the temporal and spatial characteristics of taxi passenger travel hot spots, and examine the causes of the formation of spatiotemporal characteristics.

2. Data Description

Nanjing is located in the southwest of Jiangsu Province, China. The study area selected in this paper covers the main urban area of Nanjing, including Gulou, Xuanwu, Jianye, Qinhuai, and Yuhuatai, as shown in Figure 1.

The data used in this study includes two parts of the Nanjing taxi trajectory and feature dataset. The source of Nanjing taxi trajectory data was Datatang (https://www.datatang.com), which contains data from approximately 7,800 taxis with a sampling interval of 30 seconds. Data for the same period for three consecutive years were selected: January 25–31, 2015; February 13–19, 2016; and February 2–8, 2017. The source of road network data was Tianditu (https://www.tianditu.gov.cn), a national geographic information public service platform, which contains eight types of graded roads. Considering the nature of taxi services, railways, subways, light rails, and high-speed rails were excluded. Approximately 2,400 road sections classified as national roads, provincial roads, county roads, township and village roads, and other roads were used in the analysis. The data source of points of interest was Baidu POI (http://www.data-shop.net/tag/), which includes four types of land use: land for commercial use, residential land, land for public management and public service, and land for transportation. A total of approximately 26,000 points of interest were selected.

3. Methodology

This study used ArcGIS to perform map matching and geocoding preprocessing on taxi movement trajectories and the feature data of Nanjing City. A road network geographic database and a road network topology map were created using the complex network-modeling tool NetworkX to map the road intersections. The abstraction of the road intersections is a complex network node, the corresponding road section is abstracted as an edge, and community detection is performed on the taxi pick-up and drop-off points. Based on community detection, the passenger travel activity index is constructed, and hot spot mining is realized through spatial statistics. The technical process is shown in Figure 2.

3.1. Community Detecting. First, based on the concept of a dual graph [22], the road is defined as a generalized network composed of nodes and edges.

$$\begin{aligned} G &= \left\{ \left(N_g, E_g \right) \mid 1 \le g \le n_{\text{Graph}} \right\}, \\ \left| N_g \right| &= n_{\text{node}}, \\ \left| E_g \right| &= n_{\text{edge}}, \end{aligned} \tag{1}$$

where (N_g, E_g) represents any road segment, n_{Graph} represents the total number of road segments, n_{node} represents the number of nodes included in a road segment, and n_{edge} represents the number of edges included in a road segment.

The road is abstracted into a complex network, as shown in Figure 3: (a) is the original road graph, which contains 9 road sections and 14 nodes; (b) is the corresponding original

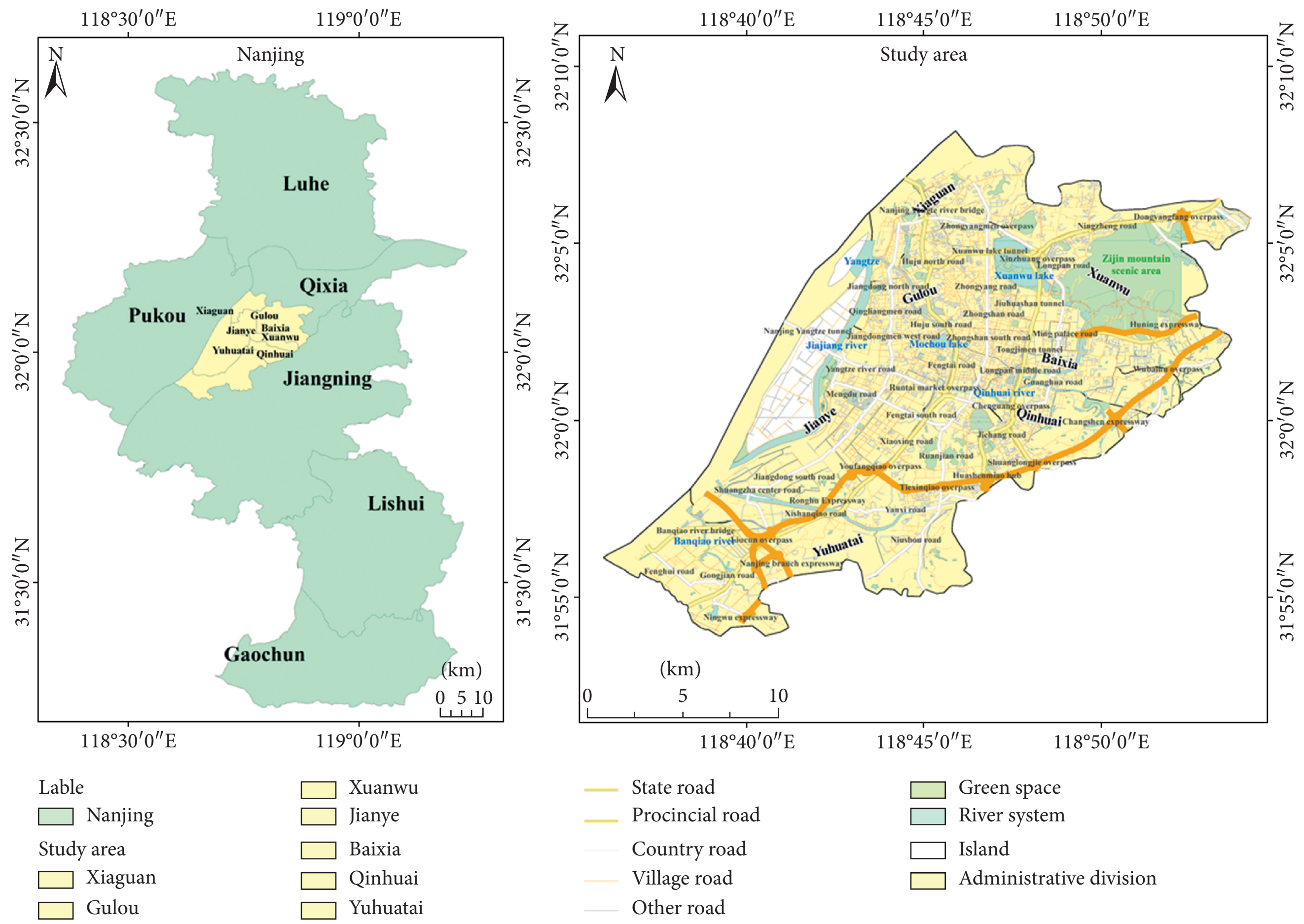


Figure 1: Study area map.

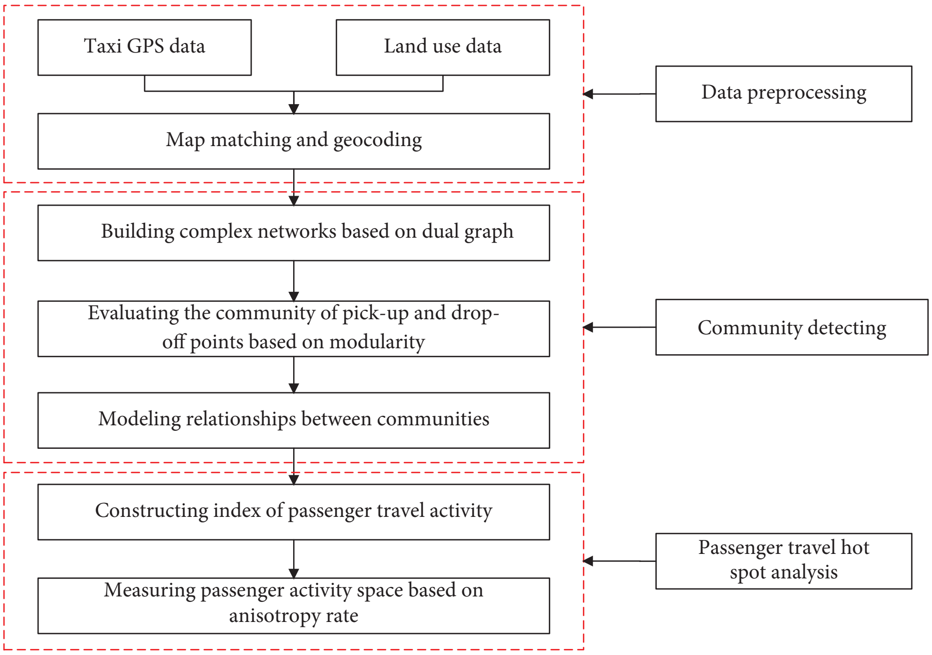

Taxi GPS data
Land use data
Map matching and geocoding
Data preprocessing
Building complex networks based on dual graph
Evaluating the community of pick-up and drop-off points based on modularity
Community detecting
Modeling relationships between communities
Constructing index of passenger travel activity
Measuring passenger activity space based on anisotropy rate
Passenger travel hot spot analysis

Figure 2: Technical flowchart.

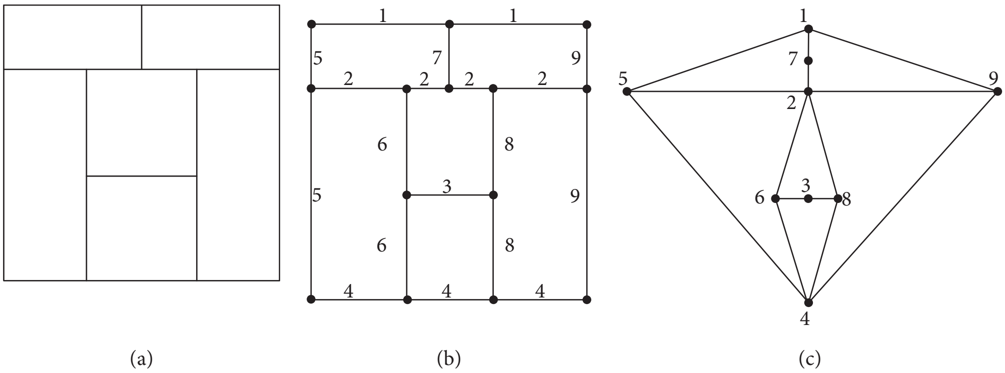


Figure 3: Diagram of a complex network of roads: (a) road network graph; (b) road network initialization graph; (c) road network dual graph.

graph, where the nodes represent road intersections and the edges represent road sections between nodes; (c) is the corresponding dual graph, in which nodes represent road sections and edges represent the intersecting relationship of road sections.

Figure 3(c) is used to abstract the road network and express intersections and road sections as nodes and edges, respectively. Before community detection, the stay points are mapped to the road network, which characterizes the geographic location using map matching technology. After abstraction, the road network not only retains the geographic location information but also characterizes the network connectivity. The connectivity of the road network changes with the degree centrality of each node in the abstract network. The more the pick-up points, the higher the exit degree of the node, and the more the drop-off points, the higher the entry degree of the node. Therefore, when the stay points change dynamically in the road network, they are aggregated into clusters according to Bayes' rule, and communities are established to represent the activities of the community's residents as hot spots.

Second, we define the stay point, corresponding to the stay point to the network, and use it as a separate atomic cluster [23].

$$S = \left\{ \left(S_i^O, S_i^D \right) \mid \quad 1 \le i \le N_S \right\}, \tag{2}$$

where (S_i^O, S_i^D) represents any pair of stay points, N_S represents the number of stay points included in the candidate dataset, S^O represents the pick-up point in a pair of stay points, and S^D represents the drop-off point in a pair of stay points.

Subsequently, the distance between the stay points is calculated according to the Euclidean metric. Taking the above stay points as an example (the same is true for the drop-off point), the two closest points are continuously merged into the same cluster, and the distance between the clusters is calculated according to the average distance measurement:

$$d_{\text{avg}}\left(C_i, C_j\right) = \frac{1}{n_i n_j} \sum_{Q_i \in C_i} \sum_{Q_j \in C_j} \left| Q_i - Q_j \right|, \tag{3}$$

where C_i and C_j are the clusters where the pick-up points S_i^O and S_j^O are located; n_i and n_j are the number of pick-up points contained in clusters C_i and C_j, respectively; and $|Q_i - Q_j|$ is the distance between clusters C_i and C_j. Q is the incremental matrix, which is the adjacency matrix that stores the nodes and edges within the cluster. Q is calculated as follows [24]:

$$Q = \frac{1}{E_g} - \frac{k_i k_j}{2E_g^2}, \tag{4}$$

where E_g is the total number of connected edges, k_i is the degree of node i, and k_j is the degree of node j.

For a network with N nodes, the execution process of the algorithm used in this study includes the following steps:

(1) *Initializing*. Treat each node as a cluster and set the increment matrix to $Q = 0$.

(2) *Merging and Updating*. Combine the two clusters C_i and C_j that have edges connected in such a way that maximizes d_{avg}, maximize d_{avg} using Bayes' rule, and update the combined cluster.

(3) *Terminating*. Continue the merging and updating process until there are no clusters that can be merged.

We then count the number of pick-up points in each cluster as the amount of information I and set the amount of information as the weight according to Bayes' rule, thereby establishing a community as

$$\begin{gathered} I = \frac{1}{n} \sum_{i=1}^{n} C_{n_i}^i, \\ P\left(\frac{n_i}{I}\right) = \frac{P\left(I/n_i\right) \times P\left(n_i\right)}{P\left(n_i\right) \times P\left(I/n_i\right) + P\left(\overline{n_i}\right) \times P\left(I/\overline{n_i}\right)}, \\ B_i = \frac{\sum_{i=1}^{n} \left(x_i, y_i\right) \times P\left(n_i/I\right)}{\sum_{i=1}^{n} P\left(n_i/I\right)}, \end{gathered} \tag{5}$$

where $C_{n_i}^i$ means that the cluster C^i in the community contains n_i pick-up points, and n is the total number of pick-up points in the community. $P(n_i/I)$ means that the pick-up points are aggregated into the pick-up point community, and (x_i, y_i) denotes the coordinates of the pick-up points. B_i is the center of the mass coordinates of community (x_B, y_B).

Finally, the community is evaluated according to the degree of modularity M:

$$M = \frac{1}{2E}\sum_{ij}\left(C_{ij} - \frac{k_i k_j}{2E}\right)\partial\left(B_i, B_j\right), \tag{6}$$

where E is the number of edges in the network. C_{ij} takes the value 0 or 1; if $C_{ij} = 1$, there is an edge between the pick-up points S_i^O and S_j^O; otherwise, there is no edge; k_i and k_j are the degrees of the pick-up points S_i^O and S_j^O; B_i and B_j are the centroids of the communities where the pick-up points S_i^O and S_j^O are located. Only when $B_i = B_j$, $\partial(B_i, B_j) = 1$. The value range of M is $[0, 1]$; the larger the value, the more obvious the community structure.

3.2. Constructing an Indicator of Passenger Travel Activity.

Based on the community detection results, the passenger travel activity point is set as $S_i = [S_i^O, S_i^D, B_i^O, B_i^D]$, where S_i^O is the coordinate of the taxi passenger pick-up point, S_i^D is the coordinate of the drop-off point, B_i^O is the centroid coordinate of the community to which the pick-up point belongs, and B_i^D is the centroid coordinate of the community to which the drop-off point belongs. Therefore, there are three situations of inclusion, intersection, and separation of the communities, in which the passenger board and drop-off points belong, as shown in Figure 4.

As shown in Figure 4, the center of the pick-up point community is O and the radius is r_O; the center of the drop-off point community is D and the radius is r_D; and the smallest circle center that contains the pick-up point community is C and the radius is R. No interaction between the pick-up point community and the drop-off point community is shown by the white area in the figure, while an interaction between the pick-up point community and the drop-off point community is shown by the shaded area in the figure.

The passenger travel activity index is a combination of outbound visit heat and arrival visit heat [25], with 1 h as the unit time for sampling and 1 km as the unit distance for calculation, defined as $A_i = [depart_i, arrive_i]$; the calculation is as follows:

$$\begin{aligned}
\text{scope}_i &= 1 - \frac{d_i}{2R_i},\\
p\left(S_i \in B_i\right) &= \begin{cases} 1, & S_i^O \notin \odot B_i^D \text{且} S_i^D \notin \odot B_i^O,\\ 0, & {}_i^O \in \odot B_i^D \text{或} S_i^D \in \odot B_i^O, \end{cases}\\
\text{depart}_i &= \sum_{S_i^O \in B_i^O}^{n_O} p\left(S_i^O \in \odot B_i^O\right) \times \text{scope}_i^O,\\
\text{arrive}_i &= \sum_{S_i^D \in B_i^D}^{n_D} p\left(S_i^D \in \odot B_i^D\right) \times \text{scope}_i^D,
\end{aligned} \tag{7}$$

where d_i is the distance between taxi passengers' pick-up and drop-off points, R_i is the smallest radius of the circle that contains the community of the pick-up point, n_O denotes all pick-up points included in the pick-up point community, and n_D denotes all drop-off points included in the drop-off point community. $scope_i$ is a probability density function that describes the distance between the pick-up point and the centroid of the community it belongs to; $depart_i$ represents the probability density estimation from the pick-up point to the pick-up point community, namely, the popularity of passenger outbound travel activities; $arrive_i$ represents the probability density estimation from the drop-off point to the drop-off point community, that is, the popularity of passenger arrival travel activities.

Given N nodes, there can be at most $N(N-1)/2$ edges, and a random network can be obtained by randomly selecting M edges from these edges. Obviously, a total of $C_{N(N-1)/2}^M$ random graphs are possible, each with the same probability. When the node's connection probability p exceeds the critical probability $p_c(N) = \ln N/N$, every random graph is connected. Therefore, a random graph Q with N nodes and connection probability $p = p(N)$ satisfies

$$\lim_{N\longrightarrow\infty} P_{N,p}(Q) = \begin{cases} 1, & p(N)/p_c(N) \longrightarrow \infty,\\ 0, & p(N)/p_c(N) \longrightarrow 0. \end{cases} \tag{8}$$

For the community of pick-up and drop-off points formed by the corresponding pick-up and drop-off points, when $p(N)/p_c(N) \longrightarrow \infty$, the random network is completely connected, forming a closed network with no isolated nodes. In other words, the pick-up point will not belong to the community of the drop-off points, and the drop-off point will not belong to the community of the pick-up points, namely, $S_i^O \notin \odot B_i^D$ and $S_i^D \notin \odot B_i^O$. When $p(N)/p_c(N) \longrightarrow 0$, the random network has a tree structure, and there are branch nodes belonging to other connected subgraphs, namely, $S_i^O \in \odot B_i^D$ or $S_i^D \in \odot B_i^O$.

Because passengers are not necessarily restricted to moving in certain pairs of communities, the standard deviation ellipse method [26] is used to measure the spatial distribution characteristics of passenger activities and the interaction of the community where the passengers travel activity points are evaluated according to the anisotropy rate, that is, an equal ellipse. The higher the anisotropy rate under the area, the more directional and purposeful the passenger activities in the community. The anisotropy rate α is calculated as follows:

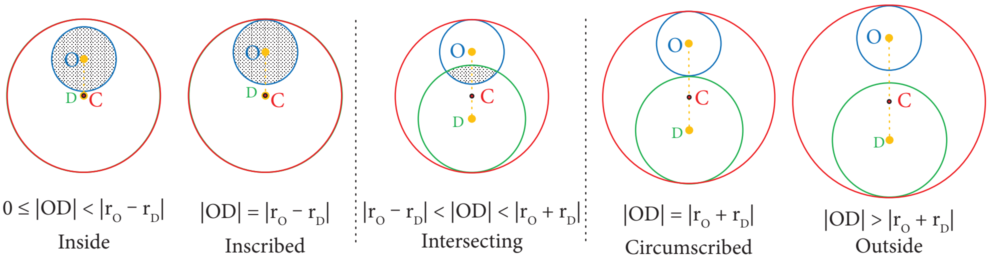
$$\alpha = \frac{\sigma_{x'} - \sigma_{y'}}{\sigma_{x'}} \times 100\%, \tag{9}$$

where $\sigma_{x'}$ is the length of the major axis of the ellipse and $\sigma_{y'}$ is the length of the minor axis of the ellipse.

4. Results and Analysis

4.1. Spatial and Temporal Characteristics of Passenger Travel Activity.

One hour was adopted as the unit time interval to summarize the passenger pick-up and drop-off points



Community inclusion
$C = IF(r_O > r_D, O, D)$
$R = MAX\ (r_O, r_D)$

Community intersecting
$C = AVERAGE(O, D)$
$R = IF(r_O > r_D, CO + r_O, CD + r_D)$

Community separation
$C = AVERAGE(O, D)$
$R = IF(r_O > r_D, CO + r_O, CD + r_D)$

FIGURE 4: Schematic diagram of three community situations. Note. The blue circle represents the taxi pick-up community, its center is O, and its radius is r_O. The green circle represents the taxi drop-off community, its center is *D*, and its radius is r_D. The red circle represents the smallest circumscribed circle that contains two communities, its center is C, and its radius is R. In the case of community inclusion, if $r_O > r_D$, then C is O; otherwise, C is D. The value of *R* is maximum in r_O and r_D. In the case of community intersection or community separation, the value of C is average of O and *D* coordinates. If $r_O > r_D$, then the value of *R* is the sum of the values of CO (distance value between C and O) and r_O; otherwise, the value of *R* is the sum of the values of CD (distance value between C and D) and r_O.

recorded in the taxi trajectory data for the weeks included in the data analysis (Section 2), as shown in Figure 5.

It can be seen from Figure 5 that the number of taxi passengers getting on and off is consistent across days of the week, and there are fluctuations at different times of the day. The daytime is higher than the nighttime, and there is a significant increase during the morning rush hour. Moreover, there is also a certain increase during the evening rush hour. Thus, taxi passenger travel show more daytime activity, less nighttime activity, and frequent activity during the morning and evening peak hours.

Taking one hour as the unit time interval, the average modularity and anisotropy rate of the communities where taxi passengers were picked up and dropped off in 2015, 2016, and 2017 are shown in Figure 6.

It can be seen from Figure 6 that, during the morning and evening peak hours, the modularity is relatively high, and the anisotropy rate curve is relatively steep. When the modularity increases, the anisotropy rate also increases. In other periods, the modularity is relatively low, and the anisotropy rate curve is relatively flat. When the modularity decreases, the anisotropy rate increases. This shows that the community structure of taxi passengers' pick-up and drop-off points becomes closer as the purpose of passengers' travel increases. For example, during morning peak hours, passengers travel mainly from home to office; during evening peak hours, passengers travel mainly from office to home; and in other periods, residents' activities are affected by differences in travel motivation, thus showing randomness.

To clearly reflect the differences in residents' travel activities at different times, the morning peak hours were 8 : 00–9:00, working hours 13 : 00–14 : 00, evening peak hours 18 : 00–19 : 00, and rest period 22 : 00–23 : 00. We can conduct community detection at the points where taxi passengers board and alight, as shown in Figure 7.

It can be observed from Figure 7 that during the period of 8 : 00–9:00, the corresponding communities of the pick-up and drop-off points are separate, and during the period of 13 : 00–14 : 00, the corresponding communities of the pick-up and drop-off points are mainly intersecting. The pick-and-drop points, as shown in Community No. 1 (Figure 7(b)), are mainly gathered in the southeast of Gulou, southwest of Xuanwu, west of Qinhuai, and northeast of Jianye area. During the period of 18 : 00–19 : 00, the corresponding communities of the pick-up and drop-off points are mainly separate. During 22 : 00–23 : 00 (Figure 7(d)), the corresponding communities of the pick-up and drop-off points are mainly inclusive. The pick-up and drop-off points shown in Community No. 1 are mainly concentrated in the northwest of Xuanwu, and the pick-up and drop-off points shown in Community No. 2 are mainly concentrated in Jianye. In the northeast, the pick-up and drop-off points shown in Community No. 5 are mainly concentrated in the northeast of Yuhuatai, and the pick-up and drop-off points shown in Community No. 6 are mainly concentrated in the middle of Gulou.

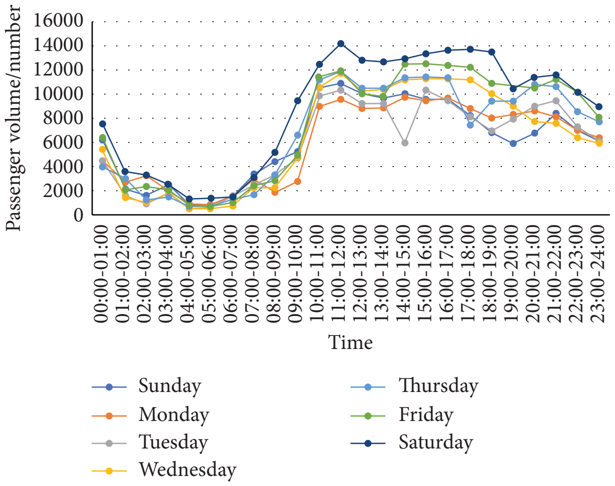


FIGURE 5: Graph of average passenger volume.

This shows that, during the same period, passenger travel activities are affected by the purpose of travel, showing the same behavioral pattern in the same community, obvious

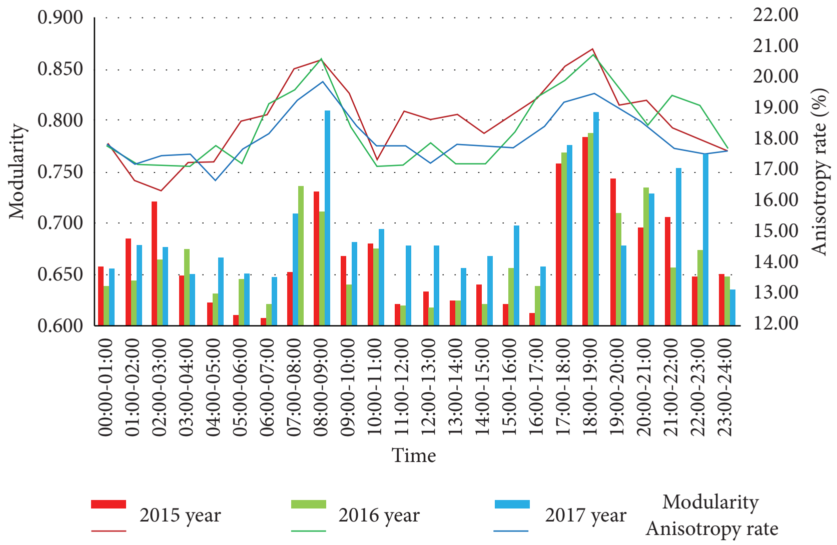
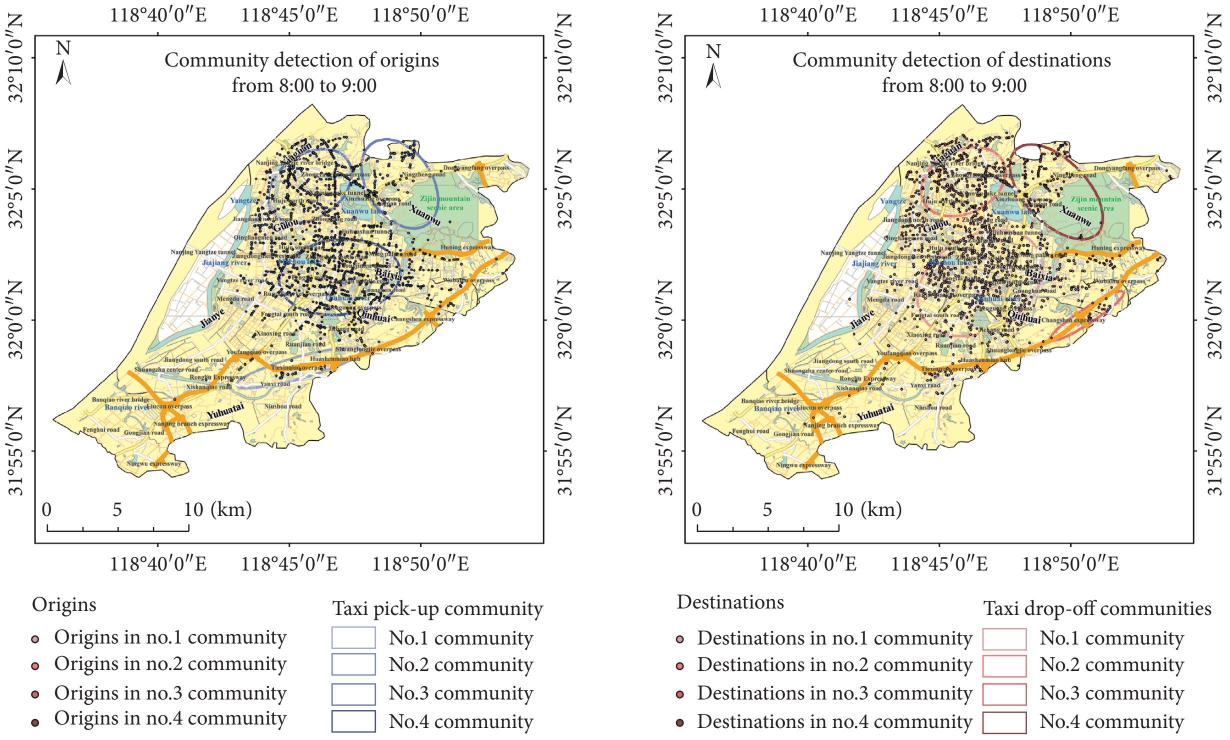
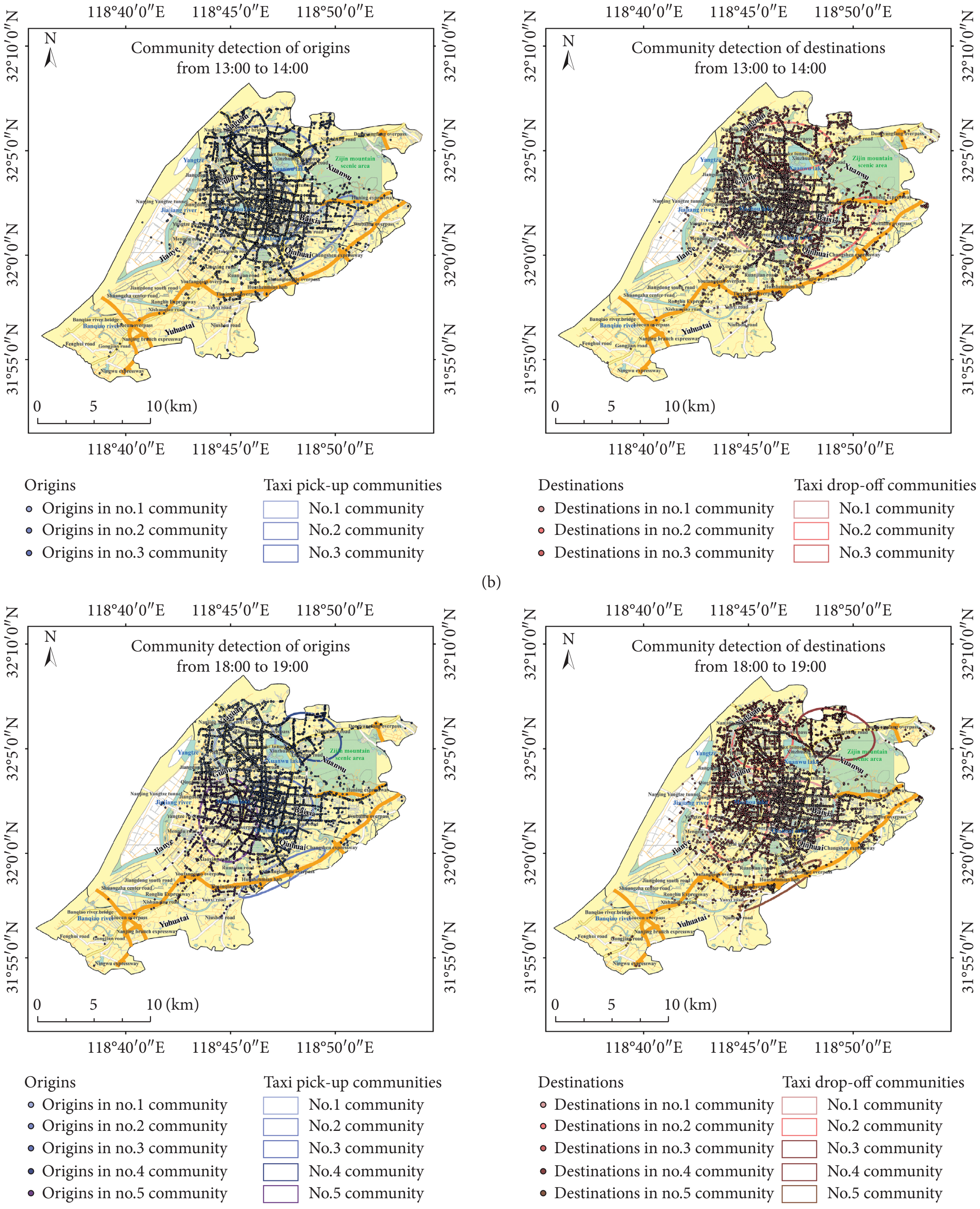


Figure 6: Graph of average community modularity and anisotropy rate.



(a)

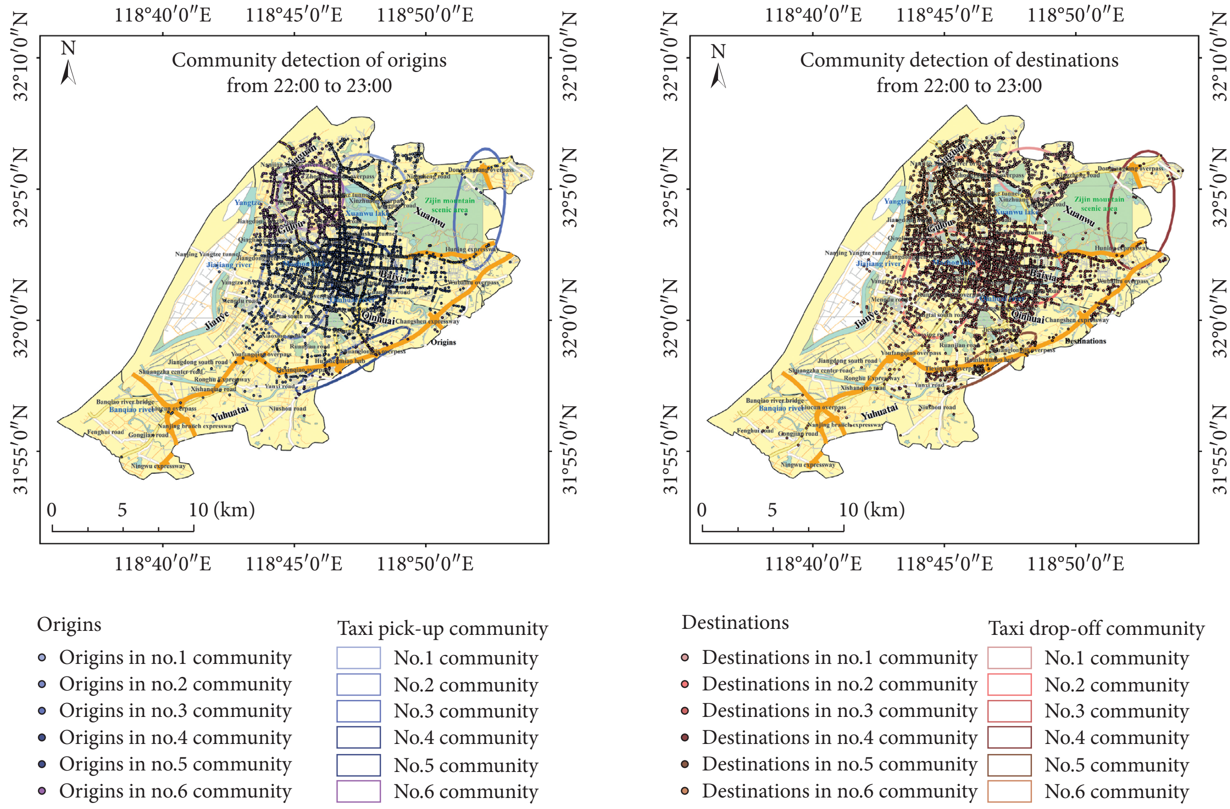
Figure 7: Continued.


Community detection of origins
from 13:00 to 14:00
Community detection of destinations
from 13:00 to 14:00
Community detection of origins
from 18:00 to 19:00
Community detection of destinations
from 18:00 to 19:00
N
118°40′0″E
118°45′0″E
118°50′0″E
32°10′0″N
32°5′0″N
32°0′0″N
31°55′0″N
0
5
10 (km)
Yangtze
Jiajiang river
Xuanwu
Zijin mountain scenic area
Baixia
Qinhuai
Jianye
Yuhuatai
Banqiao river
Origins
Origins in no.1 community
Origins in no.2 community
Origins in no.3 community
Origins in no.4 community
Origins in no.5 community
Taxi pick-up communities
No.1 community
No.2 community
No.3 community
No.4 community
No.5 community
Destinations
Destinations in no.1 community
Destinations in no.2 community
Destinations in no.3 community
Destinations in no.4 community
Destinations in no.5 community
Taxi drop-off communities
No.1 community
No.2 community
No.3 community
No.4 community
No.5 community

(b)

(c)

Figure 7: Continued.


Community detection of origins
from 22:00 to 23:00
Community detection of destinations
from 22:00 to 23:00
N
118°40′0″E
118°45′0″E
118°50′0″E
32°10′0″N
32°5′0″N
32°0′0″N
31°55′0″N
0
5
10 (km)
Origins
Destinations
Origins
Origins in no.1 community
Origins in no.2 community
Origins in no.3 community
Origins in no.4 community
Origins in no.5 community
Origins in no.6 community
Taxi pick-up community
No.1 community
No.2 community
No.3 community
No.4 community
No.5 community
No.6 community
Destinations
Destinations in no.1 community
Destinations in no.2 community
Destinations in no.3 community
Destinations in no.4 community
Destinations in no.5 community
Destinations in no.6 community
Taxi drop-off community
No.1 community
No.2 community
No.3 community
No.4 community
No.5 community
No.6 community

(d)

Figure 7: (a) Community detection results of origin and destination from 8 : 00 to 9 : 00. (b) Community detection results of origin and destination from 13 : 00 to 14 : 00. (c) Community detection results of origin and destination from 18 : 00 to 19 : 00. (d) Community detection result of origin and destination from 22 : 00 to 23 : 00.

spatial clustering, and differences between different communities.

In summary, the characteristics of the passenger travel activity time distribution show more daytime and fewer nighttime activities and frequent peak hours in the morning and evening. The characteristics of the passenger travel activity spatial distribution show concentrated urban centers and scattered peripheral areas. Affected by the purpose of travel and structure of the community, passenger travel activities behave in the same way in the same community at the same time, and there is an interaction between different communities at the same time.

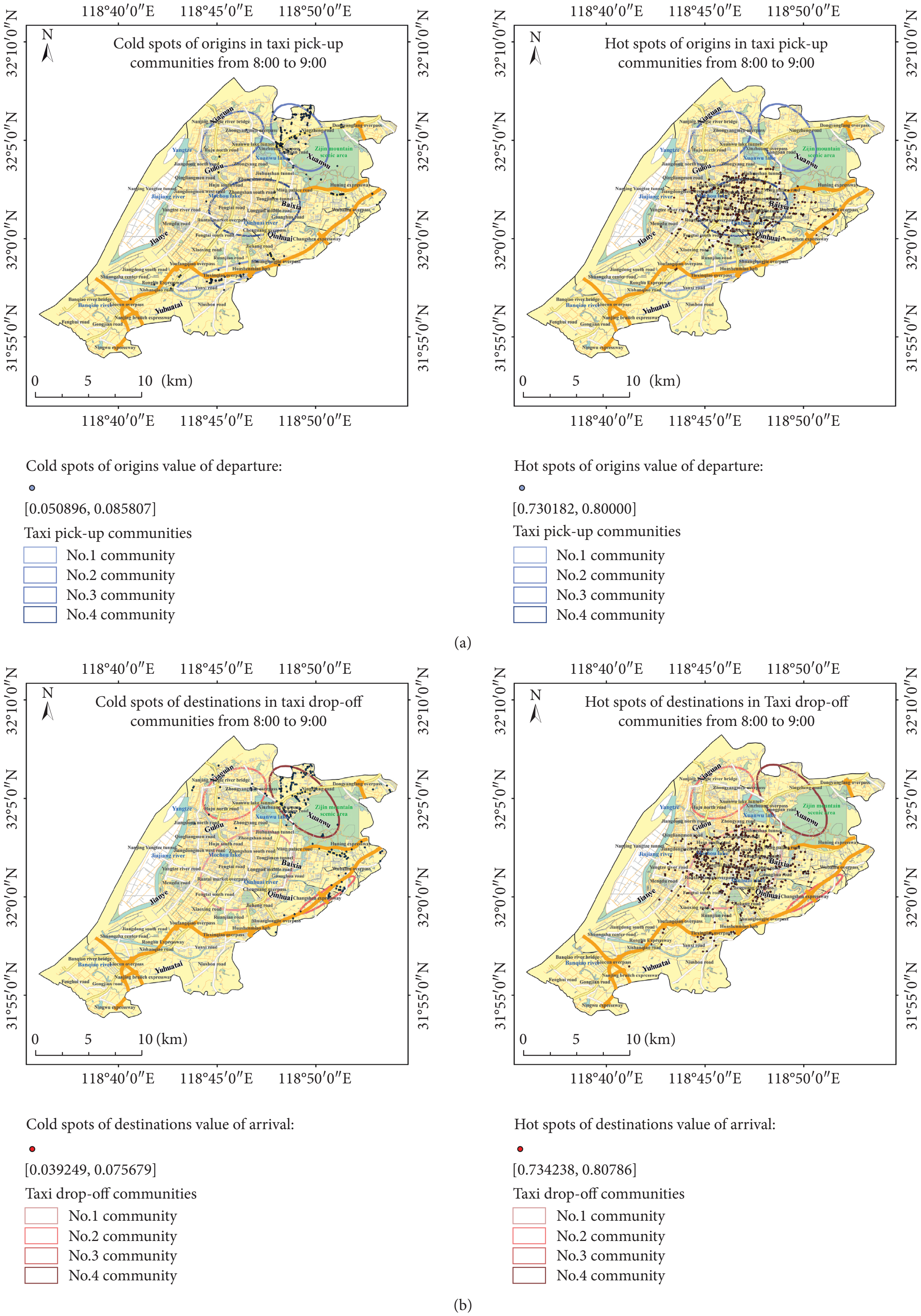
4.2. Passenger Travel Hot Spot Analysis. Taking the 8 : 00–9 : 00 time period as an example, we considered the minimum circle radius of the community, including the pick-up and drop-off points, as the aggregation distance, and the outbound visit heat and arrival visit heat of the passenger's travel activity as the indicators. The corresponding communities were divided according to the first decile, and the pick-up and drop-off points were aggregated to extract hot spots. Furthermore, the pick-up and drop-off points in the corresponding community were aggregated according to the last decile to extract cold spots, as shown in Figure 8.

It can be clearly observed from Figure 8 that the pick-up points of the decile before the visit were distributed in Community No. 4, and the drop-off points of the decile before the visit were mostly distributed in Community No. 1: in the southeast of Gulou, southwest of Xuanwu, west of Qinhuai, northeast of Jianye, and north of Yuhuatai. These are adjacent areas of the main urban centers, and the pick-up or drop-off points in the decile after the visit heat and the visit heat were randomly distributed. Therefore, it can be concluded that passenger travel hot spots were clustered or dispersed as passenger travel activity increased or decreased.

Xu et al. [19] showed that the hot spots in Nanjing have a spatial distribution characteristic of clustering from the surroundings to the center and that Moran's I value around the clustering center is negative. The hot spots of passenger travel extracted in this study are consistent with the results of the previous study, and a more obvious spatial local agglomeration can be found based on community detection of pick-up and drop-off points.

Taking the period from 8 : 00 to 9 : 00 as an example, the average visit heat and average visit heat statistics were calculated on five graded road sections: national highway, provincial highway, county highway, township and village highway, and other roads, as shown in Table 1.

It can be seen from Table 1 that the average visit popularity ordered from high to low was township and village roads, provincial roads, other roads, county roads, and national roads. The average visit popularity ordered from


Cold spots of origins in taxi pick-up communities from 8:00 to 9:00
Hot spots of origins in taxi pick-up communities from 8:00 to 9:00
Cold spots of destinations in taxi drop-off communities from 8:00 to 9:00
Hot spots of destinations in Taxi drop-off communities from 8:00 to 9:00
N
118°40′0″E
118°45′0″E
118°50′0″E
32°10′0″N
32°5′0″N
32°0′0″N
31°55′0″N
0
5
10 (km)
Cold spots of origins value of departure:
[0.050896, 0.085807]
Hot spots of origins value of departure:
[0.730182, 0.80000]
Taxi pick-up communities
No.1 community
No.2 community
No.3 community
No.4 community
Cold spots of destinations value of arrival:
[0.039249, 0.075679]
Hot spots of destinations value of arrival:
[0.734238, 0.80786]
Taxi drop-off communities
No.1 community
No.2 community
No.3 community
No.4 community
Gulou
Xuanwu
Baixia
Qinhuai
Jianye
Yuhuatai
Xiaguan

(a)

(b)

Figure 8: (a) Hot and cold spots of origins in pick-up communities from 8 : 00 to 9 : 00. (b) Hot and cold spots of destinations in drop-off communities from 8 : 00 to 9 : 00.

Table 1: Average activity of graded roads.

Graded road	Number of roads	Average value of departure	Average value of arrival
National roads	66	0.530	**0.598**
Provincial roads	84	0.648	0.474
County roads	181	0.594	0.546
Township and village roads	1191	**0.694**	0.582
Other roads	178	0.626	0.592

high to low was national roads, other roads, town and village roads, county roads, and provincial roads.

By searching the database, we found that the representative road sections with higher outbound visits on the graded roads were Yurun Street, Fengqi Road, Jiajiang Bridge, Jinsu'an Road, and Caodu Lane; representative sections with lower outbound visits were Fengwu Road, Binjiang Road, Jiangshan Street, Chuanjiang Street, and Houde Road. Representative road sections with higher arrival visits were Zhenxing Road, Shuangtang Road, Jiangshan Street, Xiaofenqiao, and Fanjiatang. Representative sections with lower arrival visits were Fengwu Road, Moxiang Road Overpass, Nanjing Yangtze River Tunnel, Lingyin Road, and Kuitou Alley. This shows that passenger travel activities are closely related to the traffic functions carried by the graded roads. The main function of expressways is to enable continuous traffic, of trunk roads is to enable transportation, of secondary trunk roads is to enable distribution traffic, and of branch roads is to enable service in local areas.

A study by Yang [20] showed that there is cascading failure and congestion in the traffic system of Nanjing. Travel conditions of people in an unbalanced road network load are affected by the coupling of sub-road networks. The results of passenger travel activity on graded roads in this study are consistent with the conclusions of that research.

Taking the 8 : 00–9:00 period as an example, based on the residents' walking considerations, the pick-up and drop-off points are the center of the circle with a radius of 300 m for coverage, covering commercial land, residential land, public management and public service land, and transportation land. The average outbound visit heat and average arrival visit heat were calculated for approximately 30 types of land use involving a total of 26,000 points of interest, as shown in Table 2.

It can be seen from Table 2 that the land use type with the highest average outbound visit heat was urban residential land, and the land use type with the lowest average outbound visit heat was commercial and financial land. The land use type with the highest average arrival visit heat was commercial and financial land. The lowest average arrival visited land use type was urban residential land.

By searching the database, the representative point of interest with higher average outbound visits was Yangzhuang Village, corresponding to Shiyang Road. The representative point of interest with lower average outbound visits was Flower Building, corresponding to Software Avenue. The representative point of interest with higher average arrival visits was Commercial Century Plaza, corresponding to the Xinjiekou commercial pedestrian area. The representative point of interest with lower average arrival visits was Sun Ye Village, corresponding to Longzang Avenue. This shows that passenger travel activities were closely related to the zoning functions carried by land use types.

A study by Jin and Xu [21] showed that the inflow and outflow on the key nodes of Nanjing's road network of different levels have an obvious hierarchical structure, and different points of interest play a certain role in the flow of tourists. The results of passenger travel activity at different points of interest in our study were consistent with the conclusions of the previous research.

In summary, urban roads contain information about the classification functions of expressways, arterial roads, secondary arterial roads, and branch roads and are affected by land use types. The pick-up and drop-off points with high passenger travel activity were concentrated near points of interest, forming hot spots. On the contrary, the pick-up and drop-off points with low passenger travel activity were concentrated near points of interest, and cold spots were formed. The hot spots of outbound visits were scattered on urban residential land, and the hot spots of arrival visits were concentrated on commercial and financial land.

5. Comparison and Discussion

5.1. Comparison. The GN algorithm [16] includes a splitting algorithm that uses the number of shortest paths passing through each edge in the network as a measurement index, and gradually deletes edges that do not belong to any community. Newman's fast algorithm [17] uses a cohesive algorithm, starting with each node occupying a community and continuously merging in the direction that maximizes the increase in modularity. Compared with the GN algorithm and the Newman fast algorithm, we use Bayes' rule to set the weight of the edge betweenness of the network, and the heap data structure to calculate the modularity; we also reduce the complexity of the algorithm and use the standard deviation ellipse to make the detected community structure clearer. For a complex network with n nodes and m connecting edges, the comparison results of the GN algorithm, Newman fast algorithm, and the algorithm in this paper are listed in Table 3.

Theoretically, if there are n communities, an $n \times n$ symmetric matrix F can be defined. The trace of the matrix (the sum of the diagonal elements of the matrix) is $Tr(F) = \sum_i f_{ii}$, which means the ratio of all edges connecting the nodes within the community to the total number of edges in the network. $Tr(F)$ value is in the range of $[0, 1]$. It is used to calculate modularity, and to a certain extent also characterizes the complexity of the network structure.

Table 2: Activities at different points of interests.

Land use types		Interests	Total	Value of departure	Value of arrival
Commercial	Retail land	Shopping malls, supermarkets, etc.	4381	0.722	0.578
	Dining land	Hotels, restaurants, etc.	5210	0.700	0.568
	Financial land	Office buildings, financial centers, etc.	462	0.276	0.648
	Other land	Banks, business halls, etc.	2284	0.738	0.590
Residential	Residential land	Apartments, villas, etc.	2161	0.790	0.288
Public	Agency land	Government agencies, etc.	810	0.674	0.574
	Education land	Schools, institutes, etc.	1376	0.614	0.508
	Medical land	Hospitals, pharmacies, etc.	1881	0.660	0.540
	Green land	Parks, gardens, etc.	74	0.554	0.450
Traffic	Street land	Parking lot, transportation station, etc.	324	0.700	0.576
	Highway land	Toll station, bus station, etc.	7206	0.308	0.444

Table 3: Method comparison.

Characteristic	GN algorithm	Newman fast algorithm	Algorithm of this paper
Algorithm complexity	$O(nm^2)$	$O(n^2)$	$O(m\log^2 n)$
Number of communities	Unknowable	Knowable	Knowable
Community structure	No overlap	Overlap	Overlap

When the network structure is abnormally chaotic, there are fewer edges connecting nodes within the community, and the value of $Tr(F)$ is minute. When the network structure is abnormally single, there are excessive number of edges connecting the nodes within the community, and the value of $Tr(F)$ is extremely large. When $Tr(F)$ value is in the range of $[0.4, 0.6]$, it can be assumed that the network structure is normal and that the value is not an abnormal value.

Therefore, another way of expressing modularity is $M = Tr(F) - F^2$, and F^2 is the modulus of matrix F^2. We compare the accuracy of community detection models using the GN algorithm, Newman fast algorithm, and the algorithm in this paper, as shown in Figure 9.

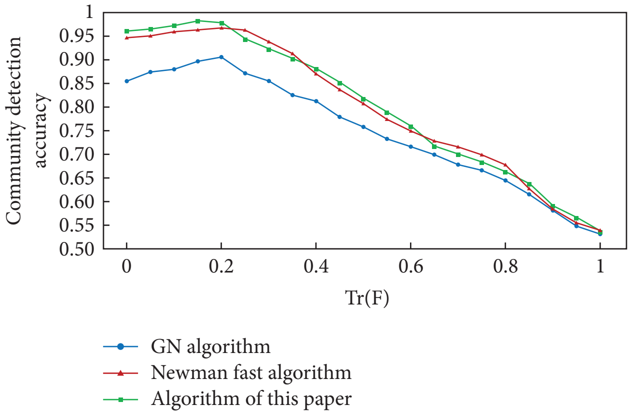


Figure 9: Algorithm performance comparison chart.

The community detection algorithm centered on the hierarchical structure is divided into split and aggregation types. The GN algorithm belongs to the split type, and the Newman fast algorithm and the algorithm proposed in this article belong to the aggregation type.

The GN algorithm gradually deletes edges that do not belong to any community (i.e., the edges connected between communities) according to the degree to which the edges do not belong to the community, until all edges are deleted. Because the edge betweenness of each connected edge needs to be recalculated every time an edge is removed, for complex network structures, the algorithm can be easily implemented by splitting it across more independent communities.

Newman's fast algorithm starts with each node occupying a community and continues to merge communities in the direction that maximizes the increase in modularity until the entire network merges into one community. Because the modularity needs to be increased every time the communities connected by edges are merged, when the network structure is simple, the execution of this algorithm will easily lead to the incorrect division of nodes.

The algorithm proposed in this paper introduces Bayes' rule and takes the amount of information as the increment of modularity, without calculating the adjacency matrix to ensure the increment of modularity. Therefore, when unconnected communities are merged, the degree of modularity remains unchanged; thus, the communities that are connected by edges and the corresponding internal nodes can be divided more accurately.

As shown in Figure 9, the abscissa is $Tr(F)$ and the ordinate represents the accuracy of community detection. The circle is the GN algorithm, the triangle is the Newman fast algorithm, and the square is the algorithm used in this study. It can be clearly observed from the figure that the accuracy of the algorithm in this study is significantly higher than that of the GN algorithm. Compared with the Newman fast algorithm, when $Tr(F)$ is $[0, 0.2]$, $[0.4, 0.6]$, and $[0.8, 1]$, the algorithm used in this study has higher accuracy. Therefore, according to Figure 9 and Table 3, the accuracy of the algorithm in this study is equivalent to that of the Newman fast algorithm, but the running time is faster, and thus the performance is better.

This demonstrates that when the network structure is abnormally single or chaotic, the community detection model using the algorithm proposed in this study can discover more complex community structures and has better interpretability for community detection results.

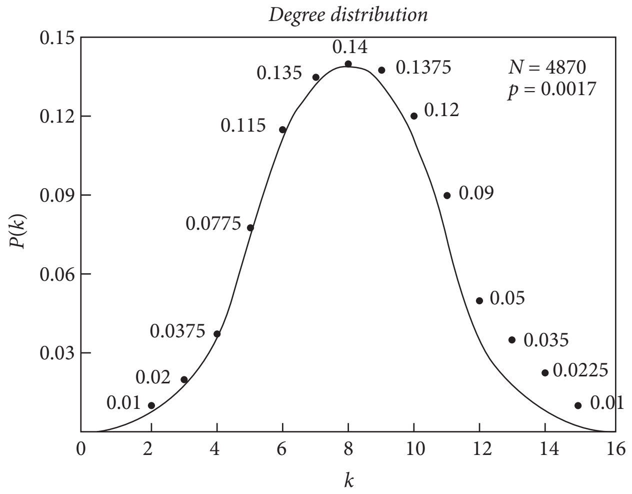


Figure 10: Degree distribution of the model.

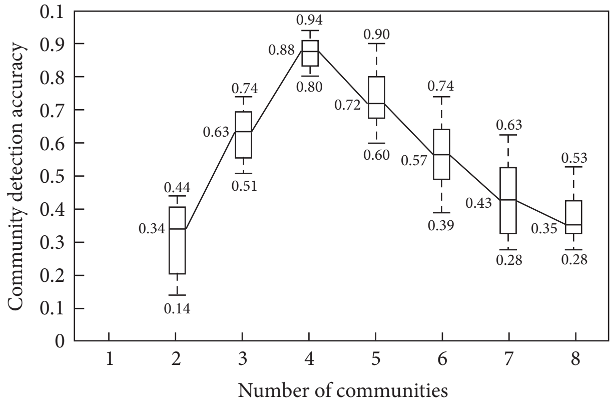


Figure 11: Box plot of model parameter sensitivity analysis.

5.2. Discussion. In order to explore the parameter sensitivity of the community detection model in this study, taking the period of 8 : 00–9:00 as an example, the degree distribution of the random network was calculated, as shown in Figure 10, and 21 simulation experiments were performed to compare the accuracy, as shown in Figure 11.

As shown in Figure 10, the abscissa represents the degree of the node, the ordinate represents the degree distribution probability, N represents the number of nodes, and p represents the connection probability of the nodes. It can be clearly observed from Figure 10 that the average degree of the node is eight, and the degree distribution follows the Poisson distribution.

As shown in Figure 11, the abscissa represents the number of communities, and the ordinate represents the accuracy of community detection. It can be clearly observed from Figure 11 that when the number of communities detected is four, the accuracy reaches its peak.

In summary, in a random network composed of 4870 key road network nodes, different communities are delineated based on taxi passengers' pick-up and drop-off points within a representative period, and the detected travel hot spots have reasonable spatial distribution characteristics.

Qin et al. [18] analyzed the intensity of node access degrees and edge weights based on the network interaction of urban hot spots, without considering the potential impact of land use on urban residents' travel decisions. This study combined the hierarchical road network and point of interest data to explore hot spots from the perspective of individual taxi passengers interacting with the community, which helped to explore the formation process of urban hot spots.

6. Conclusions

This study extracted the passenger pick-up and drop-off points from taxi movement trajectory data, constructed a taxi passenger travel activity index based on community detection, and extracted the hot spots of taxi passenger travel in the main urban area of Nanjing. The following three conclusions were drawn:

(1) The travel activities of taxi passengers showed a time distribution pattern of more daytime, less nighttime, and frequent morning and evening peak hours. Affected by the purpose of travel, the degree of community modularity and anisotropy rate of taxi passengers' pick-up and drop-off points were positively correlated during morning and evening peak hours and negatively correlated during other periods.

(2) The travel activities of taxi passengers presented a spatial distribution pattern, in which the central area of the city was concentrated and the outer areas were scattered. Affected by the structure of the community, passenger travel activities showed a consistent behavioral pattern within the community and had obvious spatial gathering characteristics. Furthermore, there was a significant interaction between different communities.

(3) The hot spots for taxi passengers' travel were scattered on urban residential land and concentrated on commercial and financial land. Affected by land use, passenger travel activity indicators were closely related to road grades and types of points of interest. Passenger travel hot spots were clustered as activity levels increased and dispersed as activity levels decreased.

Subsequent research needs to consider more sources of data, such as combining rental car trajectory data with bus trajectory data, analyzing the travel preferences of different groups of people, and further exploring the temporal and spatial patterns of urban traffic congestion by urban residents using the impact of different travel modes.

Data Availability

All data, models, and code that support the findings of this study are available from the corresponding author upon reasonable request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Authors' Contributions

Yuyu Sheng, Shuoben Bi, and Wenwu He conceived and designed the experiments; Yuyu Sheng and Ruizhang Xu performed the experiments; Yuyu Sheng, Shuoben Bi, and Ruizhuang Xu wrote the Chinese paper; Shuoben Bi and Jingjing Fan translated the paper.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grants: 41971340 and 41271410).

References

[1] Q. Q. Li and D. R. Li, "Big data GIS," *Geomatics and Information Science of Wuhan University*, vol. 39, no. 06, pp. 641–644, 2014.

[2] Y. Zheng, "Introduction to urban computing," vol. 40, no. 01, , pp. 1–13, Geomatics and Information Science of Wuhan University, 2015.

[3] Y. Liu, X. Liu, S. Gao et al., "Social sensing: a new approach to understanding our socioeconomic environments," *Annals of the Association of American Geographers*, vol. 105, no. 3, pp. 512–530, 2015.

[4] J. M. Harvey and J. W. Han, *Geographic Data Mining and Knowledge Discovery*, CRC Press, London, UK, 2009.

[5] C. X. Cheng, P. J. Shi, C. Q. Song et al., "Geographic big-data: anew opportunity for geography complexity study," *Acta Geographica Sinica*, vol. 73, no. 08, pp. 1397–1406, 2018.

[6] M. Veloso, S. Phithakkitnukoon, and C. Bento, "Urban mobility study using taxi traces," in *Proceedings of the 2011 International Workshop on Trajectory Data Mining and Analysis – TDMA'11*, pp. 23–30, ACM, New York, NY, USA, 2011.

[7] R. Ahas, A. Aasa, Y. Yuan et al., "Everyday space-time geographies: using mobile phone-based sensor data to monitor urban activity in Harbin, Paris, and Tallinn," *International Journal of Geographical Information Science*, vol. 29, no. 11, pp. 2017–2039, 2015.

[8] R. W. Scholz and Y. Lu, "Detection of dynamic activity patterns at a collective level from large-volume trajectory data," *International Journal of Geographical Information Science*, vol. 28, no. 5, pp. 946–963, 2014.

[9] J. Cui, F. Liu, D. Janssens, G. Wets, and M. Cools, "Detecting urban road network accessibility problems using taxi GPS data," *Journal of Transport Geography*, vol. 51, no. 12, pp. 147–157, 2016.

[10] C. Zhong, S. M. Arisona, X. Huang, B. Michael, and G. Schmitt, "Detecting the dynamics of urban structure through spatial network analysis," *International Journal of Geographical Information Science*, vol. 28, no. 11, pp. 2178–2199, 2014.

[11] Q. Huang, Y. Yang, Z. Yuan et al., "The temporal geographically-explicit network of public transport in Changchun City, Northeast China," *Scientific Data*, vol. 6, no. 190026, pp. 1–10, 2019.

[12] F. Guo, D. Zhang, Y. Dong et al., "Urban link travel speed dataset from a megacity road network," *Scientific Data*, vol. 6, no. 61, pp. 1–8, 2019.

[13] H. Hamedmoghadam, M. Ramezani, and M. Saberi, "Revealing latent characteristics of mobility networks with coarse-graining," *Scientific Reports*, vol. 9, no. 1, pp. 1–10, 2019.

[14] Yu Zheng, Y. Liu, J. Yuan et al., "Urban computing with taxicabs," in *Proceedings of the 13th International Conference on Ubiquitous Computing - UbiComp'11*, pp. 89–98, ACM, Beijing, China, 2011.

[15] R. L. Wu, X. Y. Zhu, and W. Guo, "Spatiotemporal distribution patterns of urban road traffic accidents," *Geomatics & Spatial Information Technology*, vol. 41, no. 07, pp. 103–106, 2018.

[16] M. Gong and M. Girvan, "Finding and evaluating community structure in networks," *Physical Review E Statistical Nonlinear & Soft Matter Physics*, vol. 69, no. 2, pp. 26113–26120, 2004.

[17] M. E. J. Newman, "Fast algorithm for detecting community structure in networks," *Physical Review E*, vol. 69, no. 6, Article ID 066133, 2004.

[18] K. Qin, Q. Zhou, Y. Q. Xu et al., "Spatial interaction network analysis of urban traffic hotspots," *Progress in Geography*, vol. 36, no. 9, pp. 1149–1157, 2017.

[19] J. Xu, S. B. Bi, Y. Zhang et al., "PSO-SVM model based analysis on traffic flow of road intersections in Nanjing," *Modern Electronics Technique*, vol. 39, no. 17, pp. 128–213, 2016.

[20] X. X. Yang, "On the Model and Congestion Performance of Nanjing Urban Public Traffic Systems," *Nanjing University of Posts and Telecommunications*, 2018.

[21] C. Jin and J. Xu, "Study on the tourists flow among external transport nodes and hotels in Nanjing," *Human Geography*, vol. 31, no. 05, pp. 55–62, 2016.

[22] L. Zhao, M. Deng, D. L. Peng et al., "Structural property analysis of urban Street networks based on complex network theory," *Geography and Geo-Information Science*, vol. 26, no. 05, pp. 11–15, 2010.

[23] G. N. Wang, *Spatial-Temporal Data Mining Based on GPS Trajectory and Geo-Tagged Photo Trajectory*, Central South University, Changsha, China, 2013.

[24] S. Z. Guo and Z. M. Lu, *The Basic Theory of Complex Network*, Science Press, Beijing, China, 2012.

[25] Y. Batty, *Study on Human Activity Space Patterns and Network Spatial Temporal Characteristics in Urban Cities Using Taxi Trajectory Data*, Wuhan University, Wuhan, China, 2016.

[26] Y. L. An, Z. F. Huang, W. D. Chen et al., "Spatial evolution of county economy in Anhui Province during 2001-2010," *Progress in Geography*, vol. 32, no. 05, pp. 831–839, 2013.

Hindawi
Journal of Advanced Transportation
Volume 2021, Article ID 6663022, 14 pages
https://doi.org/10.1155/2021/6663022

Research Article

Fitting Method of Optimal Energy-Running Time Curve Based on Train Operation Data of an Urban Rail Section

Lianbo Deng, Hongda Mei, Wenliang Zhou, and Enwei Jing

School of Traffic and Transportation Engineering, Rail Data Research and Application Key Laboratory of Hunan Province, Central South University, Changsha, Hunan 410075, China

Correspondence should be addressed to Lianbo Deng; lbdeng@csu.edu.cn

Received 13 November 2020; Revised 12 January 2021; Accepted 23 January 2021; Published 4 February 2021

Academic Editor: Xinyue Xu

Due to the complexity of the operation control of urban rail transit and diversity requirements for section running time standards, based on actual train operation data, this paper proposes a curve fitting method to find the interrelation between running time and energy consumption. According to features of the energy consumption-running time curve, the discriminant criterion of outliers is constructed to select the candidate fitting data set from the original data set. To fit the energy consumption-running time curve from two-dimensional scatter points, we propose a B-spline curve fitting method based on a genetic algorithm and the fitting method is proven to have high fitting accuracy and convergence speed. Furthermore, we propose an optimization method for the fitting curve based on dynamic adjustment of the fitting data set which is selected from the candidate fitting data set to obtain the optimal energy-running time curve. The validation of Guangzhou Metro's actual operation data shows that the energy-running time curve fitted and optimized by our method has lower energy and better continuity and smoothness and could be used for evaluation of train drivers' performance and energy consumption of train operation diagram.

1. Introduction

Under the background of developing green traffic and building energy-saving cities, urban rail transit has become one of the main means of solving the problem of urban traffic congestion due to its low energy consumption and low pollution. In most Chinese cities, urban rail transit has developed rapidly in the last few years. By the end of 2019, about 40 cities in Mainland China had opened 208 urban rail transit lines, with a total length of 6,736.2 km of which 5,180.6 km was subway lines, accounting for 76.9%. China's urban rail transit has developed by leaps and bounds, but it also faces many problems such as high energy costs and low management level. Therefore, reducing the energy consumption level of urban rail transit will help to give play to the advantages of urban rail transit and maintain its competitiveness.

Energy consumption directly related to urban rail transit is mainly generated by stations and trains. Train energy consumption includes operation energy consumption and auxiliary energy consumption such as lighting, ventilation, and air conditioning. Operation energy consumption is the main source of energy consumption of urban rail trains, which generally accounts for more than 50% of the train energy consumption and takes a large proportion of the operating expenditure. The effective reduction of operation energy consumption not only meets the requirements of developing green traffic but is also an effective way for urban rail transit enterprises to reduce costs and improve benefits. Therefore, it is of great significance to study the traction energy consumption of urban rail transit.

Generally, the operation curve of urban rail trains is optimized to minimize the operation energy consumption to obtain the optimal running time and energy consumption considering the requirements of section running time. In the trial operation stage of the train, the manufacturer presets several operation curves for selection during formal operation according to some parameters such as the vehicle performance, the horizontal and vertical section of lines, the section length, and the user requirements of the urban rail enterprise. Nowadays, there has been much research on the energy-saving operation strategy of urban rail trains. As early as 1980, Milroy [1] began to study the train operation energy consumption optimization problem. Based on the maximum principle, he proposed a short-distance energy-saving operation strategy for urban rail transit which included three stages: traction, coasting, and braking and established the optimal control model to minimize the energy consumption under a constant slope, laying the foundation for modern urban rail train optimal control theory. Khmelnitsky [2], on the basis of an analytical method, further considered the change in line slope, the difference of section speed limits, and the relationship between traction force and braking force with changes in speed and proposed a numerical algorithm for optimizing the operation strategy. Chang and Sim [3] established a multiobjective optimization control model considering comfort, punctuality, and energy consumption and used an improved genetic algorithm (GA) to calculate the switching point of coasting mode so as to achieve the effect of saving energy by reasonably increasing the coasting time. Ke et al. [4] optimized the section operation strategy of urban rail transit by using the maximum-minimum ant colony cloning optimization algorithm and proved that their proposed algorithm had a higher computational efficiency than other intelligent algorithms. Liu et al. [5] used the maximum principle to optimize the energy-saving operation strategy of the train and then used a numerical algorithm to solve the switching point of the train operation mode, also achieving good results. Villalba et al. [6] proposed an optimization model based on the train speed relationship and set a speed limit for trains traveling between stations to minimize energy consumption, achieving a 19% reduction in energy consumption compared to current levels. Bocharnikov et al. [7] designed a fitness function with variable weightings which was used to identify optimal train trajectories by running a series of simulations in parallel with a genetic algorithm search method and optimized traction energy consumption during a single-train journey by the optimal train trajectories.

However, the optimization of the operation curve is a complex optimization problem for which it is difficult to obtain the optimal solution. Meanwhile, operation curves preset by the manufacturer are limited and unable to cover all running time requirements of daily train operation. Therefore, aiming at the theoretical limitations of optimization of operation energy consumption at a certain running time, this paper attempts to study the relationship between section running time and reasonable energy consumption from a data-driven perspective by using the abundant train operation data formed in the operation process of urban rail enterprises.

In order to establish the relationship between section running time and optimal energy consumption, the data fitting method is very suitable. The fitting curve can visually show the changing trend of discrete data and has a wide range of applications in engineering practice. The selection of the fitting data set directly affects the accuracy and effect of the final fitting curve, so it is very important to choose a suitable fitting data set. M. Rza Mashinchi et al. [8] designed the granularity box regression method based on border regression to preprocess the data set containing outliers, eliminated outliers deviating from the fitting curve, and then conducted linear regression analysis on the data. Hossein Hassain et al. [9] demonstrated the importance of eliminating noise from the data set in plant growth curve fitting and proposed that using singular spectrum analysis to process data can effectively eliminate noise. Sanpeng Zheng et al. [10] improved the classical moving least squares method and could automatically identify outliers from the discrete data set to reduce the influence of outliers on the fitting curve through a weight function and to ensure the fitting effect. Ping Chen et al. [11] proposed a Gibbs sampling algorithm to detect additive outliers and patches of outliers in bilinear time series models based on the Bayesian view and demonstrated the efficacy of detection and estimation by Monte Carlo methods. Galvez et al. [12] applied the firefly algorithm, a powerful metaheuristic nature-inspired algorithm, to compute the approximating explicit B-spline curve for a given set of noisy data points. Trejo-Caballero et al. [13] proposed a linear combination of radial basis functions (RBFs) to tackle the curve fitting problem with a set of data points including noises.

Considering the influence of the passenger loading rates on the correlation between the running time and the energy consumption, we construct the data set of train operation based on a given loading-rate standard. The train operation data in this paper refer to the operation information of each train in the research section and operation direction, including the section operation curve, the section running time, and the corresponding energy consumption. Ignoring specific details such as the operation speed, acceleration, operation mode, and other parameters of the operation curve, we construct energy consumption-running time data points (E-T points) by taking the section running time as the abscissa and the energy consumption as the ordinate to study the change laws between running time and energy consumption and obtaining the optimal energy consumption-running time curve (the optimal E-T curve).

The optimal E-T curve shows the lowest operation energy consumption in different section running times. Meanwhile, the corresponding train operation curves of E-T points can provide abundant running curve support for the operation of the train under different running times. And due to the optimality of each point of the curve, the reasonableness of the operation strategy adopted by a train in the section can be evaluated accordingly by the comparison between the actual operation energy consumption and the optimal operation energy consumption. What is more, with the optimal E-T curve of each section in the train operation diagram, the optimal energy consumption of the entire operation diagram can also be calculated to evaluate the energy consumption level of the existing operation diagram and make up the optimal energy consumption timetable.

To obtain the optimal E-T curve, we first construct the discriminant criterion of the energy consumption level of the train operation data to eliminate the obvious unreasonable data with high energy consumption from the original train operation data set and obtain a candidate set of fitting data points after the preliminary screening to improve the accuracy of curve fitting. Secondly, based on the candidate fitting data set, a B-spline fitting method is adopted. Finally, based on the feedback of the fitting results, we establish an optimization method to improve the quality of the fitting curve by the dynamic adjustment of the fitting data set.

2. Candidate Fitting Data Set

The section operation curve of urban rail trains generally includes several parts such as "maximum traction-cruising (or coasting)-maximum braking." However, minor changes in operating strategy and control parameters will cause changes in the curve, running time, and energy consumption, resulting in a diversity of section operation data of trains. Even the same running time may correspond to several different train operation curves, as shown in Figure 1.

As described above, the train operation strategy (the operation curve) of the train operation data and the running time under the strategy are defined as E-T points. Because of the diversity of train operation data in rail sections, E-T points are relatively unordered and scattered in the coordinate system. In order to study the relationship between section running time and energy consumption and to obtain the optimal E-T curve, it is necessary to construct the discriminant criterion of outliers of E-T points, remove unreasonable energy consumption data from the original data set, and filter out a better data set to ensure the fitting effect of the E-T curve.

Although there may be a one-to-many relationship between section running time and the train operation curve, each running time should have the unique optimal energy consumption [14, 15]. Figure 2 shows that optimal energy consumption and running time have a negative correlation and the uniqueness of the optimal energy consumption in a certain running time. Thus, the outliers of E-T points can be removed according to the interrelation of the train operation data.

Taking the four data points in Figure 3 as an example, according to the approximate inverse relation between optimal energy consumption and running time, the optimal energy consumption EB at running time TB must be less than EA because TB > TA, so point B should be removed. Similarly, compared with point C, point D needs to be removed.

Setting the original data set of E-T points as $P = \{p_i(T_i, E_i), i = 1, 2, 3, \}$, according to the above analysis, the optimal data set $P^* = \{p_i^*(T_i^*, E_i^*), \quad i = 1, 2, 3, \ldots\}$ should satisfy the following law: $\forall p_1^*(T_1^*, E_1^*), p_2^*(T_2^*, E_2^*) \in P^*$, if $T_2^* \geq T_1^*$, then $E_2^*/T_2^* \leq E_1^*/T_{21}^*$, $E_2^* \leq E_1^*$, and generally, $\forall p_1^*(T_1^*, E_1^*)$, $p_2^*(T_2^*, E_2^*) \in P^*$, if $T_2^* > T_1^*$, then $E_2^*/T_2^* < E_1^*/T_{21}^*$, $E_2^* < E_1^*$.

Based on the features of the optimal data set, the discriminant criterion of outliers of the original data set is constructed as follows:

$\forall p_1 T_1,\ E_1,\ p_2 T_2,\ E_2 \in P$, if $E_1 = E_2, T_1 > T_2$, then p_1 is worse than p_2. $\forall p_1(T_1, E_1), p_2(T_2, E_2) \in P$, if $T_1 = T_2, E_1 > E_2$, then p_1 is worse than p_2. $\forall p_1(T_1, E_1), p_2(T_2, E_2) \in P$, if $T_1 \geq T_2, E_1 > E_2$ or $(E_1/T_1) > (E_2/T_2)$, then p_1 is worse than p_2.

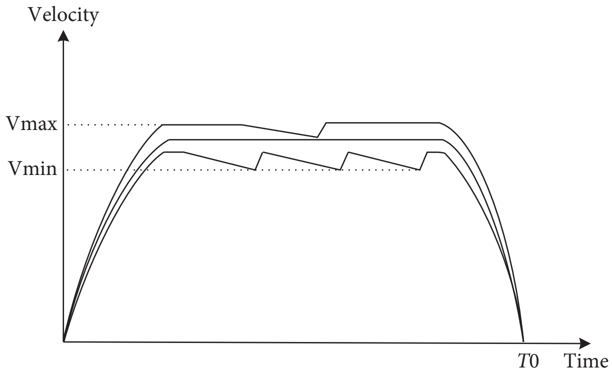


Figure 1: Multiple operation curves can correspond to the same running time.

The above rules can simply distinguish and eliminate some obvious bad points, but there may be some points of poor quality that cannot be eliminated. Therefore, we define the inferiority to measure the quality of these points, and the greater the inferiority, the worse the energy consumption of one point. $\forall p(T, E) \in P$, the neighboring point with less running time of point p is $p_1(T_1, E_1), T_1 < T$, and $\overrightarrow{p_1'}$ is the direction vector of the tangent at point p_1 on the E-T curve, so the inferiority of point p is defined as follows:

$$\delta(p) = \frac{\delta\left(\overrightarrow{p_1'}, \overrightarrow{p_1 p}\right)}{\delta\left(\overrightarrow{p_1'}, 0\right)}, \tag{1}$$

where $\delta(\overrightarrow{p_1'}, \overrightarrow{p_1 p})$ is the angle between vector $\overrightarrow{p_1'}$ and $\overrightarrow{p_1 p}$, $\delta(\overrightarrow{p_1'}, 0)$ is the angle between vector $\overrightarrow{p_1'}$ and the horizontal axis, and $\delta(p) \in (0, 1]$.

Based on the discriminant criterion of outliers, we can preliminarily remove outliers and points with unreasonable energy consumption from the original data set $P = \{p_i(T_i, E_i), i = 1, 2, 3, \ldots\}$ and obtain the candidate fitting data set $P^s = \{p_i^s(T_i^s, E_i^s), i = 1, 2, 3, \ldots\}$. The specific steps of the algorithm are as follows:

Step 1. Order all E-T points in the original data set in ascending order according to the running time. $|P|$ is the size of set P, and the initial number of elements in set P is Num.

Step 2. Remove outliers based on the discriminant criteria. Set $i = 0$, and compare the energy consumption of p_i and p_{i+1}, where $p_i, p_{i+1} \in P$. If $T_{i+1} = T_i, E_{i+1} > E_i$, or $T_{i+1} > T_i, E_{i+1} \geq E_i$, then $P = P - \{p_{i+1}\}, |P| = |P| - 1$. Set $i = i + 1$.

Step 3. If $i < \text{Num} - 1$, return to Step 2; otherwise, set $P^s = P\ d$ and take P^s as the candidate fitting data set.

In this algorithm, we obtain the candidate fitting data set by the discriminant criterion of outliers, and the definition of inferiority is used in the dynamic adjustment of the fitting data set later.

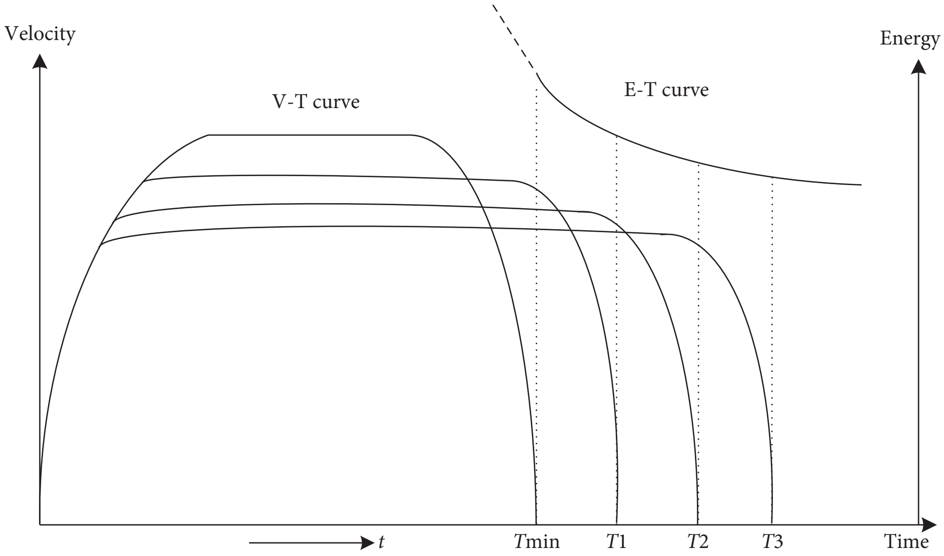


Figure 2: The interrelation between section running time and energy consumption.

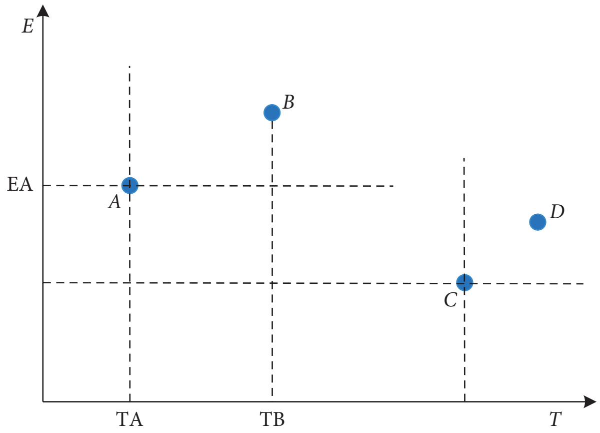


Figure 3: Removal of points with high energy consumption.

3. A B-Spline Curve Fitting Method Based on a Genetic Algorithm for the Optimal E-T Curve

For the selected train operation data set, the running time and energy consumption of the data constitute scattered points that could be fitted as an E-T curve. B-spline has the powerful function of expressing and designing free-form curves and surfaces and is one of the most popular mainstream methods for the mathematical description of shapes. So the B-spline curve can be used to fit a set of two-dimensional data points based on the train operation data [16, 17].

3.1. B-Spline Curve Fitting Method for the Selected Data Set. An ordered data set $Q = q_i\{(T_i, E_i), i = 1, \dots, m\}$ is selected from the candidate fitting data set P^s $(Q \subseteq P^s)$, and the parameter vector of Q is $T = \{t_i, i = 1, \dots, m\}$. The mathematical definition of the B-spline curve over the knot vector $U = \{u_0 = \cdots = u_k \le u_{k+1} \le \cdots \le u_n \le u_{n+1} = \cdots = u_{n+k+1}\}$ is shown as follows:

$$B(t) = \sum_{j=0}^{n} P_j N_{j,k}(t), \qquad (2)$$

where k is the degree of curve, $P_j\,(j = 0, 1, \dots, n)$ are control points, and $N_{j,k}(t)\,(j = 0, 1, \dots, n)$ are the B-spline basis functions. Basis functions are calculated using the following equations:

$$N_{j,0}(t) = \begin{cases} 1, & u_j \le t < u_{j+1}, \\ 0, & \text{otherwise}, \end{cases} \qquad (3)$$

$$N_{j,k}(t) = \frac{t - u_j}{u_{j+k} - u_j} N_{j,k-1}(u) + \frac{u_{j+k+1} - u}{u_{j+k+1} - u_{j+1}} N_{j+1,k-1}(u). \qquad (4)$$

If necessary, the convention $(0/0) = 0$ in equation (4) is applied. When the data set Q falls on the B-spline curve, $\forall q_i \in Q$ should be satisfied:

$$q_i = B(t_i) = \sum_{j=0}^{n} P_j N_{j,k}(t_i), \quad i = 1, \dots, m, \qquad (5)$$

which is written in matrix form as follows:

$$Q = NP, \qquad (6)$$

where $\mathbf{Q}$ is the data set matrix and $\mathbf{N}$ is the B-spline basis function matrix that could be calculated by the parameter vector T and the knot vector U.

The B-spline curve should go through the start point and end point and then

$$\begin{aligned} Q_0 &= B(t_0) = P_0, \\ Q_1 &= B(t_1) = P_1. \end{aligned} \qquad (7)$$

Aiming at the minimum square sum of error (SSE) between the fitting data points and actual data points, the objective function can be expressed as

$$f=\sum_{i=1}^{m-1}\left[Q_i-B(t_i)\right]^2=\sum_{i=1}^{m-1}\left[R_i-\sum_{j=1}^{n-1}P_jN_{j,k}(t_i)\right]^{2'},\quad(8)$$

where

$$R_i=Q_i-Q_0N_{0,k}(t_i)-Q_mN_{m,k}(t_i),\quad i=1,2,\ldots,m-1.\quad(9)$$

According to the least squares principle, calculate the partial derivative of control points $P_l\,(l=1,2,\ldots,n-1)$ as follows:

$$\frac{\partial f}{\partial D_l}=\sum_{i=1}^{m-1}\left[-2R_iN_{l,k}(t_i)+2N_{l,k}(t_i)\sum_{j=1}^{n}P_jN_{j,k}(t_i)\right],\quad(10)$$

and then

$$-\sum_{i=1}^{m-1}R_iN_{l,k}(t_i)+\sum_{i=1}^{m-1}\sum_{j=1}^{n-1}D_jN_{j,k}(t_i)N_{l,k}(t_i)=0,\quad(11)$$

$$\sum_{i=1}^{m-1}\left(\sum_{j=1}^{n-1}N_{j,k}(t_i)N_{l,k}(t_i)\right)D_j=\sum_{i=1}^{m-1}R_iN_{l,k}(t_i).\quad(12)$$

Transform equation (12) into matrix form and then

$$\left(\mathbf{N}^T\mathbf{N}\right)D=R,\quad(13)$$

where

$$\mathbf{N}=\begin{bmatrix}N_{1,k}(t_1)&\ldots&N_{n-1,k}(t_1)\\\vdots&\ddots&\vdots\\N_{1,k}(t_{m-1})&\ldots&N_{n-1,k}(t_{m-1})\end{bmatrix},$$

$$\mathbf{R}=\begin{bmatrix}N_{1,k}(t_1)R_1+&\cdots&+N_{1,k}(t_{m-1})R_{m-1}\\\vdots&\ddots&\vdots\\N_{n-1,k}(t_1)R_1+&\cdots&N_{n-1,k}(t_{m-1})R_{m-1}\end{bmatrix},\quad(14)$$

$$\mathbf{D}=\begin{bmatrix}D_1\\\vdots\\D_{\mathbf{n}-1}\end{bmatrix}.$$

The control point matrix $\mathbf{P}$ can be calculated approximately as follows:

$$P=\left(\mathbf{N}^T\mathbf{N}\right)^{-1}\mathbf{N}^TQ.\quad(15)$$

Furthermore, the mathematical expression of the fitting curve Q^c can be obtained as

$$Q^c=N\left(\mathbf{N}^T\mathbf{N}\right)^{-1}\mathbf{N}^TQ.\quad(16)$$

Letting $q_i^C\in Q^c$ be the point on the fitting curve corresponding to parameter t_i, and the sum of squares of the least squares error SSE is calculated as

$$\mathrm{SSE}=\sum_{i=1}^{m}\left|q_i-q_i^C\right|^2.\quad(17)$$

According to the above theory, the key of the B-spline curve fitting method is to find out the parameter vector T and the knot vector U of the fitting data set. In previous researches, the knot vector is always fixed, and then, the parameter vector c be selected by methods such as uniform parameterization, the Gauss–Newton approach, and centripetal model parameterization. Alternatively, the parameter vector is first determined by the cumulative chord length parameterization, and then, the knot vector is calculated [18, 19]. However, the accuracy of the fitting curve obtained by these methods is not satisfying, and it is difficult to obtain the optimal fitting curve. Therefore, in this paper, we combine the GA and B-spline curve fitting method to solve the nonlinear problem by changing the parameter vector and node vector simultaneously [20]. Meanwhile, considering the internal relationship between the parameter vector and the knot vector, when the parameter vector and the number of control points are determined, the appropriate knot vectors can be directly calculated by the average ordered parameter method [21, 22]. In this way, each adjacent knot interval corresponds to at least one data point that ensures that the fitting curve has high fidelity. Therefore, the fitting problem of the B-spline curve is transformed into the problem of using a GA to find out the optimal parameter vector and control points without coding the parameter vector and knot vector at the same time, which reduces the complexity of the algorithm.

3.2. Genetic Algorithm Design in B-Spline Curve Fitting

3.2.1. Notation. All the relevant notations used in the genetic algorithm are listed in Table 1.

3.2.2. Selection of the Initial Population and Chromosome Coding. The initial population of the genetic algorithm is generally generated randomly. According to the features of the B-spline curve, we generate the initial population of size N randomly in this paper. Coding methods usually include binary coding and real coding. In order to reflect the increasing characteristic of the parameter vector in the B-spline curve more intuitively, real coding is adopted in this paper. The chromosome of each individual in the population is coded as an $(\mathrm{m}+1)$-dimensional increasing real vector in the following equation:

$$\{G_1,G_2,\ldots,G_m,G_{m+1}\},\quad(18)$$

where $G_1=0, G_2=1$, and $G_{m+1}\in(0,1]$.

The top m genes of the chromosome represent the parameter vector corresponding to the fitting data, which are increasing and randomly selected within the interval $[0,1]$. While the $(m+1)th$ gene of the chromosome represents the number of control points, the number cannot be less than four because the degree of the B-spline curve is three. So we select G_{m+1} randomly within the interval $[4,m]$.

3.2.3. The Fitness Function. The fitness function is directly related to the quality of the final result and the optimization efficiency of the genetic algorithm. In order to obtain a fitting

Table 1: Notation in genetic algorithm.

	Parameters
N	Population size
ρ_c	Crossover probability
ρ_{c1}, ρ_{c2}	Parameters of the crossover probability
ρ_u	Mutation probability
ρ_{u1}, ρ_{u2}	Parameters of the mutation probability
f_{mean}	Average fitness
$f_{\max}$	Maximum fitness
$f_{\min}$	Minimum fitness
H	Set of parent chromosomes $H = \{h_i \mid i = 1, 2, \ldots, N\}$
C_c	Chromosomes produced by crossover
C_u	Chromosomes produced by mutation
C	Set of offspring chromosomes $C = C_c + C_u$
$\{G_1, G_2, \ldots, G_m, G_{m+1}\}$	Individual chromosome gene

curve with the least squares error with as few control points as possible, we design the fitness function as

$$\text{fitness} = \frac{1}{1 + \text{SSE} + \lambda \times n}, \tag{19}$$

where SSE is the least squares error and λ is the weight factor of control points.

λ affects the number of control points of the fitting curve. Generally, the more control points there are, the higher the precision of curve fitting, but it is easy to fall into "over-fitting" if there are too many control points.

3.2.4. Crossover. Common crossover methods include single-point crossover, two-point crossover, uniform crossover, and linear crossover. In this paper, the top m genes of individual chromosomes represent the parameter vector, so the new chromosome obtained after crossover should also maintain the increasing feature of parameter vectors, and thus, linear crossover is more suitable.

Two individuals are randomly selected from the set of parent chromosomes, random number r is generated within the interval $[0, 1]$, and the linear crossover is operated if $r < \rho_c$ as follows:

$$\begin{cases} h_1' = sh_1 + (1 - s)h_2, \\ h_2' = (1 - s)h_1 + sh_2, \end{cases} \tag{20}$$

where s is a random number in the interval $[0, 1]$ and h_1' and h_2' are the new chromosomes after crossover. In this way, the top m genes of the new chromosome keep increasing, and we also need to round the $(m + 1)$th gene of the chromosome to an integer because G_{m+1} is the number of control points.

Meanwhile, based on the theory of inheritance of superiority, we set the crossover probability of good chromosomes with a higher fitness to be larger than that of chromosomes with lower fitness so as to ensure that superior genes can be passed on to their offspring and improve the optimization ability of the algorithm. The dynamic crossover probability ρ_c is calculated by the fitness of the population as follows:

$$\rho_c = \begin{cases} \dfrac{\rho_{c1} - (\rho_{c1} - \rho_{c2})(f_{\text{mean}} - f_2)}{(f_{\text{mean}} - f_{\min})}, & f_1 \leq f_{\text{mean}}, \\[2ex] \dfrac{\rho_{c1} - p_{c2}(f_{\max} - f_2)}{(f_{\max} - f_{\min})}, & f_2 \leq f_{\text{mean}} < f_1, \\[2ex] \rho_{c1}, & f_2 > f_{\text{mean}}, \end{cases} \tag{21}$$

where f_1 is the larger fitness value of the two chromosomes to be crossed while f_2 is the smaller fitness.

3.2.5. Mutation. According to the coding method and the characteristics of chromosomal genes, we adapt the single-point mutation method, which means every gene on the chromosome of every individual in the population may mutate.

Let $h_1 = \{G_1, G_2, \ldots, G_m, G_{m+1}\}$ be one chromosome from the set of parent chromosomes. For each gene G_i of h_1, we generate a random number r within the interval $[0, 1]$ and operate single-point mutation if $r < \rho_u$ as

$$G_i' = \begin{cases} 0, & i = 1, \\ sG_{i+1}, & i = 2, \\ G_{i-1} + s(G_{i+1} - G_{i-1}), & 2 < i < m, \\ 1, & i = m, \\ \text{Round}(G_i + s), & i = m + 1, \end{cases} \tag{22}$$

where s is a random number in the interval $[0, 1]$, G_i' is the new chromosome after mutation, and $\text{Round}(G_i + s)$ when $i = m + 1$ means that G_{m+1} is rounded to an integer.

Also based on the theory of inheritance of superiority, in order to preserve the superior genes, we set the good chromosomes with high fitness to mutate with low probability, while the chromosomes with low fitness mutate with high probability to seek new optimization directions. The dynamic mutation probability ρ_u is calculated by the fitness of the population as

$$\rho_u = \begin{cases} \dfrac{\rho_{u1}\left(f_{\max} - f\right)}{\left(f_{\max} - f_{\text{mean}}\right)}, & f \ge f_{\text{mean}}, \\ \\ \rho_{u2}, & f < f_{\text{mean}}, \end{cases} \tag{23}$$

where f is the fitness of the chromosomes to be mutated.

3.2.6. Selection. The study shows that the convergence of the GA mostly lies in the selection operator, which may be roulette selection, the expected value selection method, and the sorting selection method. We adapt the classic roulette selection in this paper. We first calculate the fitness of the set of parent chromosomes and the set of offspring chromosomes generated by crossover and mutation, respectively, and then determine the selection probability of the individual chromosome based on its corresponding fitness ratio in all chromosomes, and finally, select the new generation population according to the selection probability.

3.2.7. Algorithm Flow

Step 1. Input the degree of the B-spline curve, the fitting data set Q, the size of the population N, the maximum iteration number $t_{\max}$, and the fitting precision ε. After the initial population is randomly generated, we set iteration number $t = 1$.

Step 2. Calculate the fitness parameters including $f_{\text{mean}}, f_{\max}$, and $f_{\min}$.

Step 3. Select a pair of chromosomes from the set of parent chromosomes randomly, calculate the crossover probability ρ_c, generate new chromosomes by the linear crossover operation, and add the new chromosomes into C_c.

Step 4. Select a chromosome from the set of parent chromosomes in turn, calculate the mutation probability ρ_u, operate single-point mutation on every gene of the chromosome, and add the mutated chromosome into C_u.

Step 5. Calculate the fitness of the set of offspring chromosomes C and select a new generation of size N from the set $H$and C by the roulette selection method.

Step 6. If $t > t_{\max}$ or $SSE \le \varepsilon$, the algorithm is terminated. Otherwise, set $t = t + 1$ and repeat steps 2 to 5.

The algorithm flowchart is shown in Figure 4.

4. Fitting Curve Optimization Based on Dynamic Adjustment of the Fitting Data Set

4.1. Optimization Model of the E-T Fitting Curve. We could preliminarily screen the original data set by the discriminant criterion of outliers described in Section 2 and obtain the candidate fitting data set. However, some points with large inferiority in the candidate fitting data set may make the curve fitting effect poor and are therefore not suitable for inclusion in the fitting data set. For example, as shown in Figure 5, points B and D could be eliminated by the discriminant criterion of outliers so the candidate fitting data set includes points $A, E, F, C,$ and G, but points $E$and F have relatively large inferiority compared with other points and are bound to affect the curve trend, which means the energy consumption of the fitting curve is not optimal. Therefore, it is important to select the optimal fitting data set from the candidate fitting data set to obtain the optimal E-T curve, so we propose an optimization model for the E-T fitting curve calculated by the fitting method in Section 3.

Each point on the optimal E-T curve fitted by the B-spline curve fitting method should satisfy the following:

$$B\left(T_i\right) \le E_{p_i^s}, \quad \forall p_i^s\left(T_i, E_i\right) \in P^s, \tag{24}$$

where $B(T_i)$ is the ordinate value of the point on the fitting curve when its abscissa value is T_i.

Based on the approximate inverse relation between the optimal energy consumption and running time, the optimal E-T fitting curve should satisfy the monotonicity and continuity of the first and second derivatives, expressed mathematically in equations (25) and (26), respectively:

$$B'\left(T_i\right) \le 0, B'\left(T_i\right) < B'\left(T_{i+1}\right), \quad \forall q_i\left(T_i, E_i\right) \in Q, \tag{25}$$

$$B''\left(T_i\right) \ge 0\, B''\left(T_i\right) > B''\left(T_{i+1}\right), \quad \forall q_i\left(T_i, E_i\right) \in Q. \tag{26}$$

At the same time, considering the feasibility of fitting, the number of points in the fitting data set should be guaranteed, shown as follows:

$$|Q| \ge 3. \tag{27}$$

On the basis of satisfying the above constraints, as many points as possible must be included. Therefore, the objective function of the optimization of the E-T fitting curve is put forward as follows:

$$\max |Q| - \lambda_1 \text{SSE}, \tag{28}$$

where λ_1 is the penalty factor of the least squares error SSE.

In summary, the optimization model of the E-T fitting curve is as follows:

$$\begin{cases} \max & |Q| - \lambda_1 \text{SSE}, \\ & B\left(T_i\right) \le E_{p_i^s}, \ \forall p_i^s\left(T_i, E_i\right) \in P^s, \\ \text{s.t.} & B'\left(T_i\right) \le 0, B'\left(T_i\right) < B''\left(T_{i+1}\right), \ \forall q_i\left(T_i, E_i\right) \in Q, \\ & B''\left(T_i\right) \ge 0\, B''\left(T_i\right) > B''\left(T_{i+1}\right), \ \forall q_i\left(T_i, E_i\right) \in Q, \\ & |Q| \ge 3. \end{cases} \tag{29}$$

4.2. Optimization Algorithm Design Based on Dynamic Adjustment of the Fitting Data Set. Firstly, the fitting data set Q is selected from the candidate data set P^s randomly, and according to the E-T curve fitted by the B-spline curve fitting method based on the GA, we constantly adjust the fitting data set to improve the fitting results. Due to the fact that the fitting precision ε is a tiny number and $\text{SSE} < \varepsilon$, SSE can be

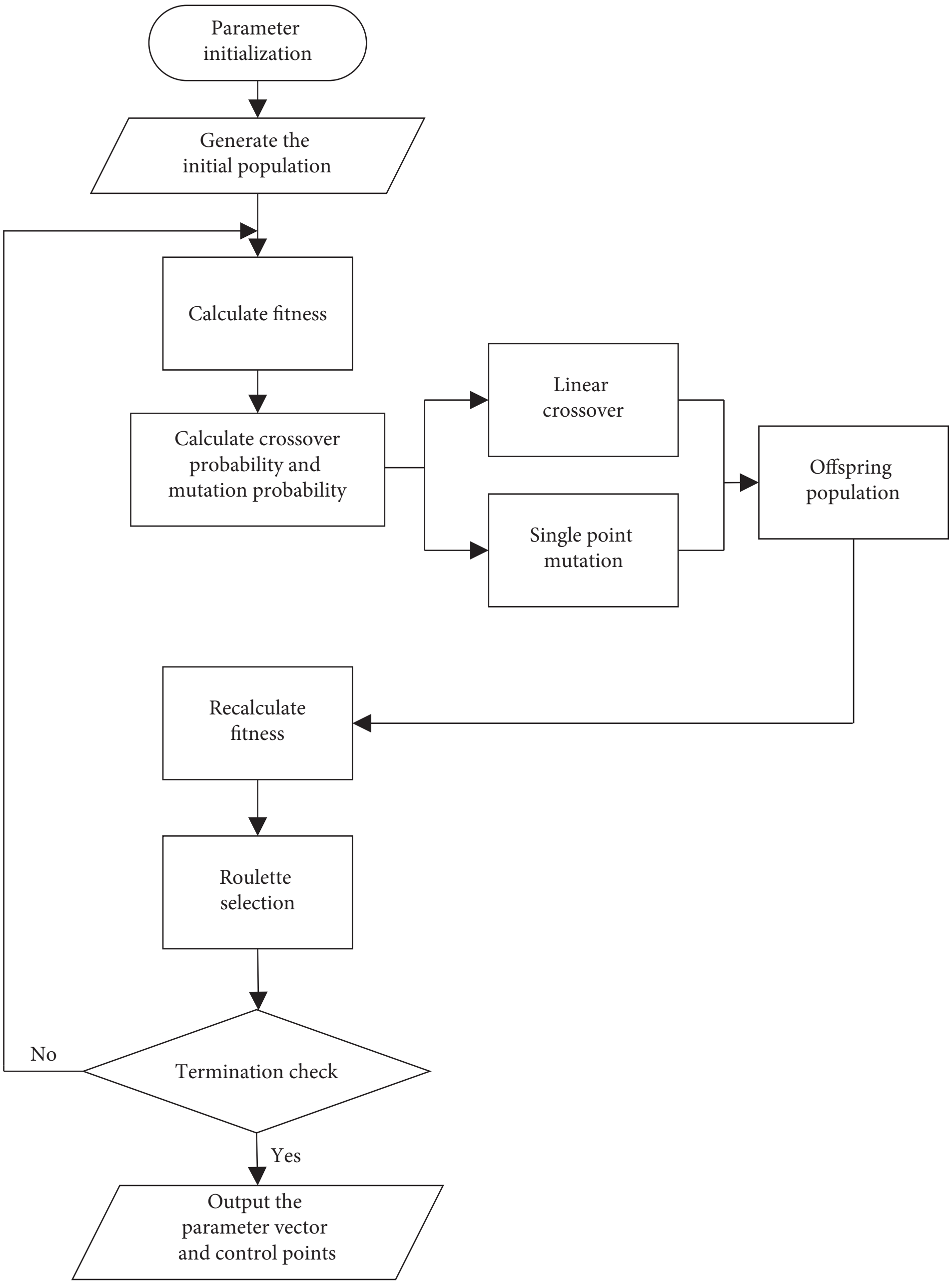


FIGURE 4: Algorithm flowchart.

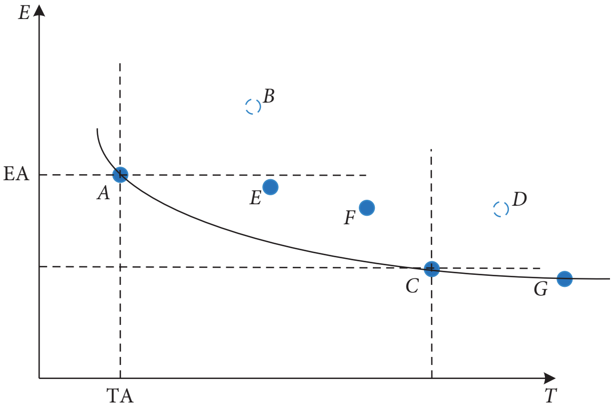


FIGURE 5: Elimination of points with larger inferiority.

regarded as zero in the objective function. The penalty factors of the constraints in equations (24), (25), and (26) are added to the objective function in the following equation:

$$\max|Q| - \lambda_2 \sum_{\substack{p_i^s(T_i,E_i)\in P^s-Q \\ B(T_i)-E_{p_i^s}>0}} \left(B(T_i) - E_{p_i^s}\right)^2 - \lambda_3 \sum_{\substack{q_i(T_i,E_i)\in Q \\ B'(T_i)>B\prime(T_{i+1}) \\ B''(T_i)<0}} (\delta(T_i))^2 - \lambda_4 \sum_{\substack{q_i(T_i,E_i)\in Q \\ B''(T_i)<0}} (B''(T_i))^2, \tag{30}$$

where λ_2, λ_3, and λ_4 are the penalty factors, and generally, $\lambda_4 < \lambda_3 < \lambda_2$ because constraint in equation (24) relates to the optimization of energy consumption, while the constraints in equations (25) and (26) relate to the smoothness of the fitting curve.

We design the optimization algorithm on the basis of the theory of tabu search. We define bidirectional tabu lists including the tabu list Q^+ ($Q^+ \in P^s - Q$) for adding points to the fitting data set and the tabu list Q^- ($Q^- \in Q$) for removing points from the fitting data set. For $q \in Q^+$ or $q \in Q^-$, $\eta(q)$ is the tabu step size of point q and η_{max} is the maximum tabu step size.

The steps to add points to the fitting data set Q are as follows:

Step 1. Calculate the data set $Q_0^+ = \{p_i^s(T_i, E_i) | p_i^s \in P^s - Q, B(T_i) - E_{p_i^s} > 0\}$ that does not meet the constraint in equation (24).

Step 2. Select the point to be added. If $Q_0^+ = \varnothing$, do not add points to the fitting data set Q. Otherwise, if $Q_0^+ \neq \varnothing, Q_0^+ - Q^+ \neq \varnothing$, select point $p \in Q_0^+ - Q^+$ by the roulette selection method and add point p into data set Q. The selection probability is calculated as follows:

$$\rho^+(p) = \frac{\left[B(T_i) - E_{p_i^s}\right]^2}{\sum_{p\in Q_0^+-Q^+}\left[B(T_i) - E_{p_i^s}\right]^2}. \tag{31}$$

If $Q_0^+ \neq \varnothing, Q_0^+ - Q^+ = \varnothing$, select one point randomly from the data set PTS $= \{p | p \in Q^+, \eta(p) = 1\}$ and add this point into the data set Q.

Step 3. Update the bidirectional tabu lists. As for the new point p added into data set Q, set $Q^- = Q^- \cup \{p\}$ and $\eta(p) = \eta_{\max}$. For $\forall q \in Q^+$, set $\eta(q) = \eta(q) - 1$, and if $\eta(q) = 0$, set $Q^+ = Q^+ - \{q\}$.

The steps to remove points from the fitting data set Q are as follows:

Step 1. Calculate the data set $Q_0^- = \{p(T_i, E_i) | p \in Q, B'(T_i) > B'(T_{i+1}), B'(T_i) < 0\}$ that does not meet the constraints in equations (25) and (26).

Step 2. Select the point to be removed. If $Q_0^- = \varnothing$, do not remove points from the fitting data set Q. Otherwise, if $Q_0^- \neq \varnothing, Q_0^- - Q^- \neq \varnothing$, select point $p \in Q_0^- - Q^-$ by the roulette selection method and remove point p from data set Q. The selection probability is calculated as

$$\rho^-(p) = \frac{[\delta(p)]^2}{\sum_{p\in Q_0^- - Q^-}[\delta(p)]^2}. \tag{32}$$

If $Q_0^- \neq \varnothing, Q_0^- - Q^- = \varnothing$, select one point randomly from data set MTS $= \{p | p \in Q^-, \eta(p) = 1\}$ and remove this point from data s Q.

Step 3. Update the bidirectional tabu lists. As for the point p removed from data set Q, set $Q^+ = Q^+ \cup \{p\}$ and $\eta(p) = \eta_{max}$. For $\forall q \in Q^-$, set $\eta(q) = \eta(q) - 1$, and if $\eta(q) = 0$, set $Q^- = Q^- - \{q\}$.

The adjustment of the fitting data set Q terminates when $Q_0^+ = \varnothing$ and $Q_0^- = \varnothing$. The algorithm flowchart is shown in Figure 6.

5. Experimental Examples

The train operation data samples in this paper are mainly composed of actual train operation data and simulated data, and the section running time range is 80–120 s. Generally, the data samples are based on the actual operation data; however, in the actual train operation process, the value range of section running time is relatively limited (82–85 s). We simulate some data samples as supplements by the software named "Urban rail transit train traction calculation and operation diagram energy consumption evaluation" which is used by the operation department of Guangzhou Metro. For each train operation data, we take the unit distance (such as 0.1 m) as the calculation step, and the energy consumption is calculated by accumulating the power (the traction force multiplies the distance) at all distance steps. We extracted a bunch of E-T points to form the original data set, as shown in Figure 7.

The energy consumption of most data points shown in Figure 7 is within a reasonable range; however, there are also some points whose energy consumption value is obviously too high. Based on the discriminant criterion of outliers, we preliminarily remove outliers and points with unreasonable energy consumption from the original data set, and all remaining points after filtering constitute the candidate fitting data set in Figure 8.

Although the overall trend of the candidate data set conforms to the characteristics of the E-T curve, it can be seen from the partially enlarged view that some points are unordered and have great inferiority that will definitely affect

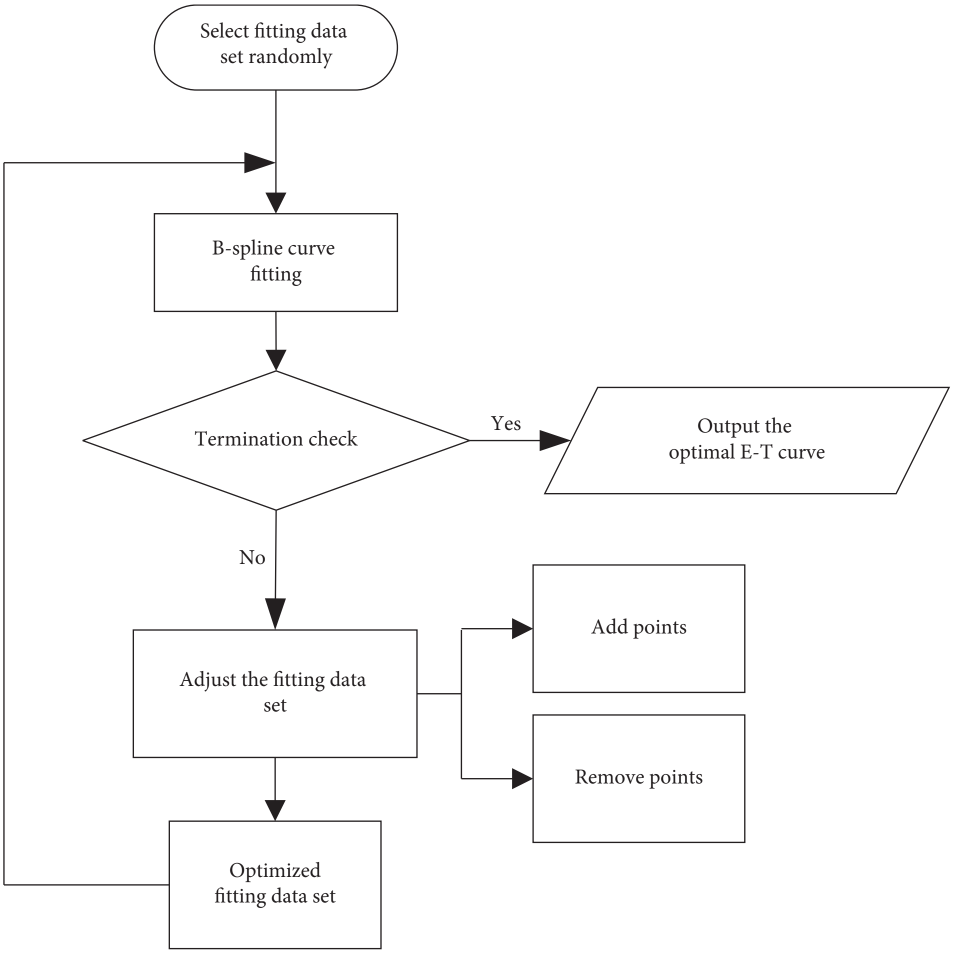

Select fitting data set randomly
B-spline curve fitting
Termination check
Yes
Output the optimal E-T curve
No
Adjust the fitting data set
Add points
Remove points
Optimized fitting data set

Figure 6: Optimization algorithm flowchart.

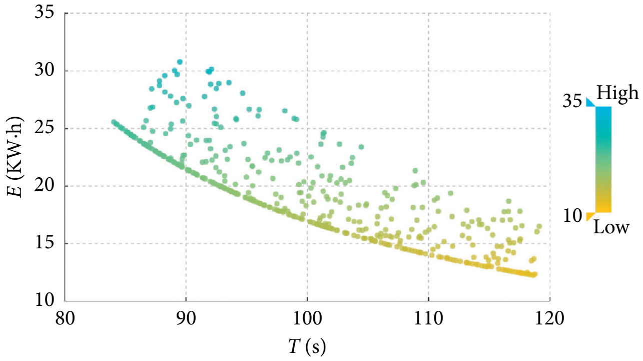


Figure 7: Scatter chart of the original data set.

the fitting effect of the E-T curve and could not ensure optimal energy consumption of the curve. Therefore, we have to select the optimal fitting data set from the candidate data set by the optimization model of the fitting curve proposed in Section 4, and the model parameter settings are shown in Table 2.

We randomly select 68 points for the fitting data set Q. From the candidate data set P^s, and as the number of iterations increases, data set Q is constantly adjusted and eventually includes 82 E-T points when $Q_0^+ = \varnothing$ and $Q_0^- = \varnothing$. The results are shown in Figures 9 and 10.

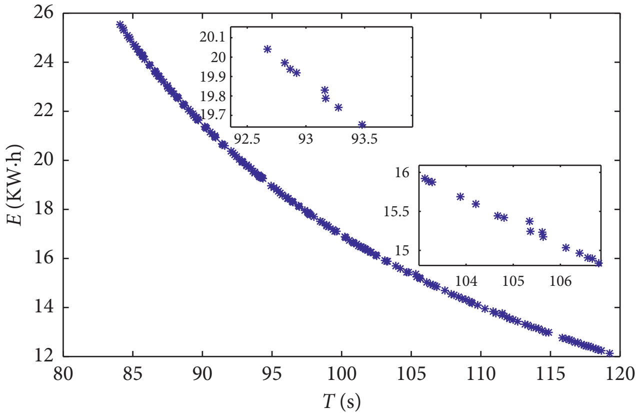


Figure 8: Scatter chart of the candidate data set.

In Figure 9, although the value of the objective function fluctuates, the overall trend is to increase with the number of iterations, and the data set Q_0^+ that does not meet the constraint in equation (17) drops significantly and finally drops to zero, indicating that the energy consumption of the fitting curve points basically reaches the optimal value. Because we have already removed some outliers from the original data set, the number of points in the data set Q_0^- that does not meet the

Table 2: Parameter values in the model.

Parameter	Value
Degree of B-spline curve k	3
Initial population size N	40
Parameters of the crossover probability ρ_{c1}, ρ_{c2}	$\rho_{c1} = 0.9, \rho_{c2} = 0.2$
Parameters of the mutation probability ρ_{u1}, ρ_{u2}	$\rho_{u1} = 0.3, \rho_{u2} = 0.8$
Fitting precision ε	0.001
Maximum tabu step size $\eta_{\max}$	3

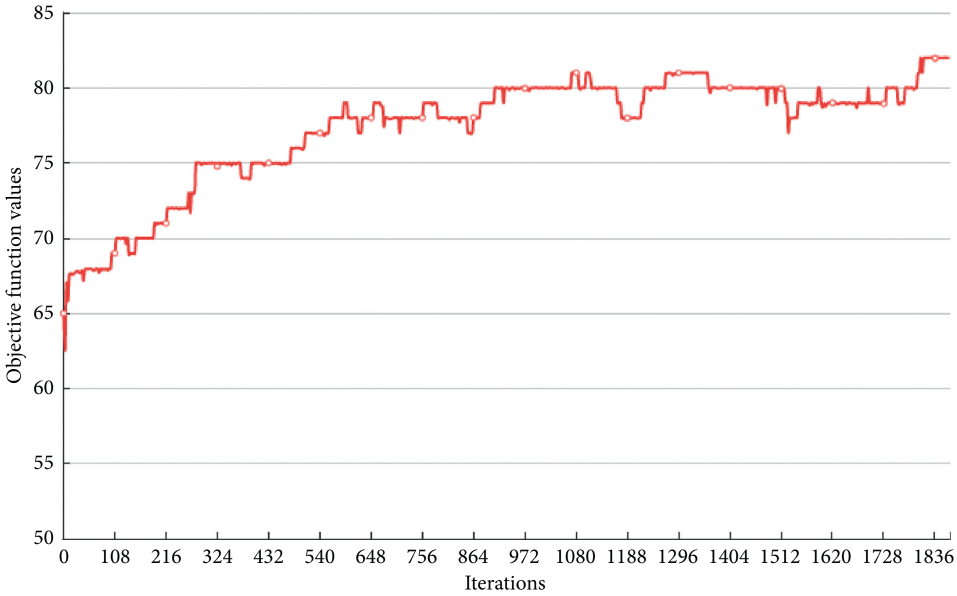


Figure 9: Change trend of the value of the objective function.

constraints in equations (18) and (19) stays low all the time and eventually decreases to zero as well, which ensures that the fitting curve has good continuity and smoothness.

The optimal energy consumption curve fitted by the optimized fitting data set using the B-spline curve fitting method based on the GA is shown in Figure 11. Figure 11(a) shows the changing trend of the least squares error and the average fitness of the population, and Figure 11(b) is the optimal E-T fitting curve.

The least squares error of the fitting curve decreases rapidly, and the average fitness of the population increases as the number of generations increases. In the 166th generation, the algorithm is terminated when the least squares error SSE = 0.00943, which meets the requirements of fitting accuracy, and the average fitness of the population also stabilizes. This proves that the B-spline fitting method based on the GA proposed in this paper has strong optimization ability, high fitting accuracy, and high convergence speed.

Meanwhile, the comparison between the optimized fitting curve and the original fitting curve fitted by the candidate data set is shown in Figure 12.

The original fitting curve is fitted by the candidate fitting data set, while the optimized fitting curve is fitted by the fitting data set optimized and adjusted by the optimization algorithm in Section 4. Although the overall trends of the two curves are similar, in a partially enlarged view, the optimized fitting curve is smoother after eliminating some points with great inferiority which have an influence on the curve, and by contrast, the optimized fitting curve is below the original fitting curve on the whole, which means that the goal of energy consumption optimization has been achieved well. Taking 0.1s as the time interval, we, respectively, select 353 sample points with the same running time from the original fitting curve and the optimal fitting curve and all sample points are from 84.1 s to 119.2 s. By accumulating the difference ratio of all sample points, the energy consumption of the optimal fitting curve is 0.69 KW h less than that of the original fitting curve, and the maximum energy consumption difference among all sample points is 0.16%. From the calculation, the energy consumption of the optimized fitting curve is lower than that of the original fitting curve and the optimal fitting data set has fewer fitting data points, which proves that the optimization method of the fitting curve proposed in this paper can select the optimal fitting data set from a large number of original data points and obtain the optimal E-T curve.

The optimal E-T curve can reflect the lowest operation energy consumption under different section running times, so each data point on the optimal E-T curve corresponds to the optimal operation curves, such as the velocity-distance (V-T) curve and time-distance (T-S) curve. In Figure 13, we

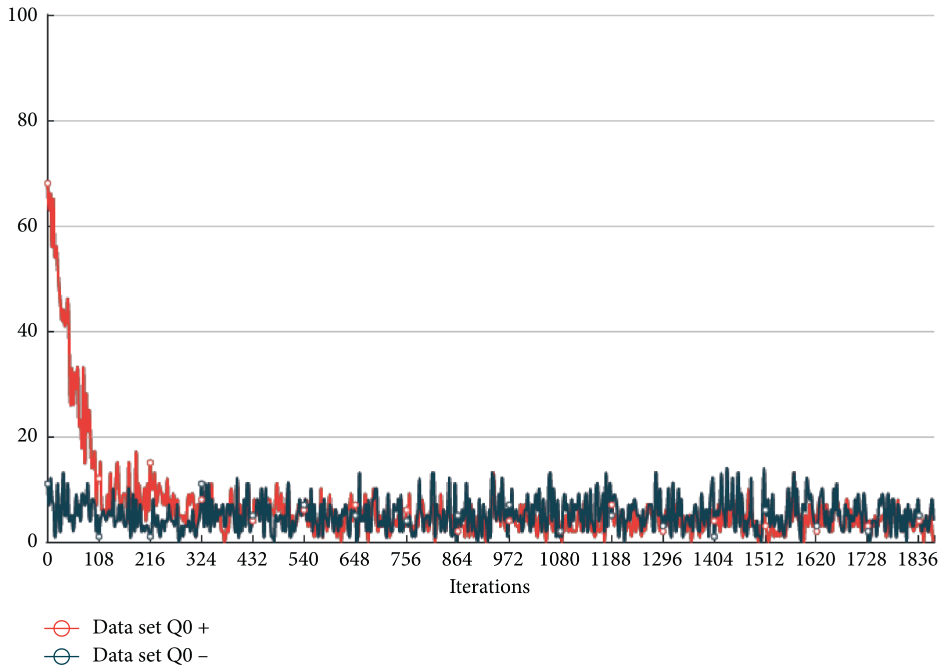
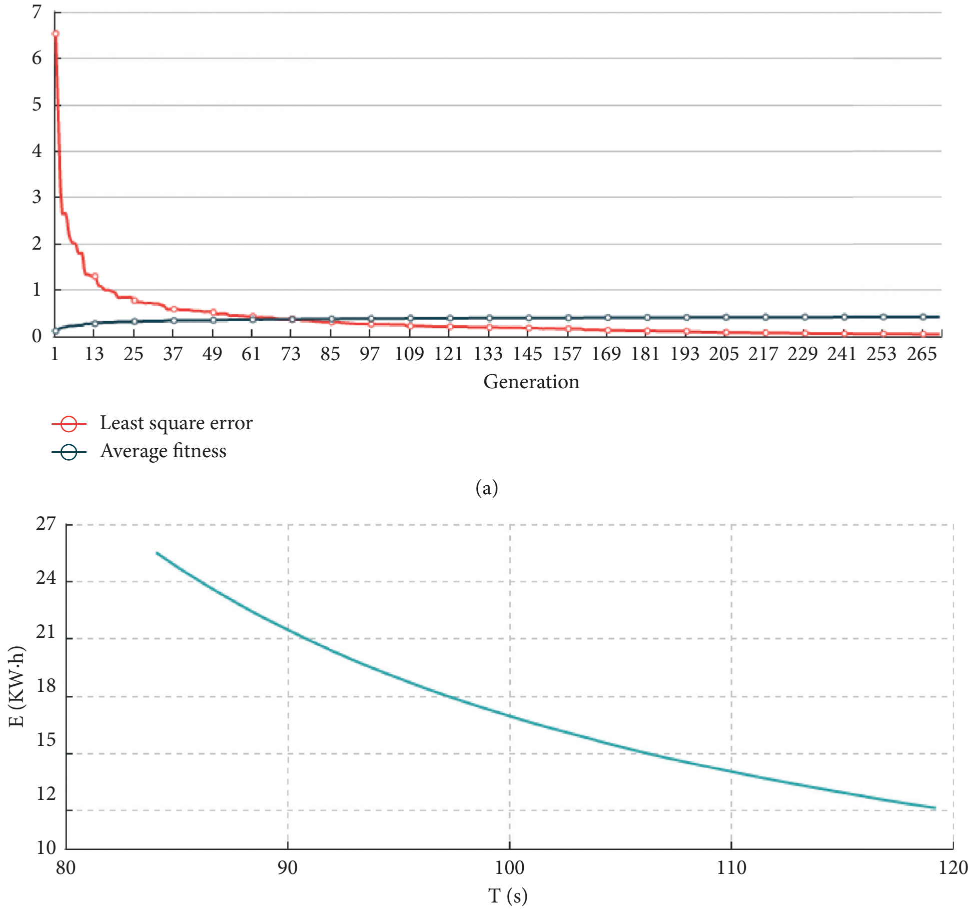


Figure 10: Change trend of data sets that do not meet the constraints.



(a)

(b)

Figure 11: B-spline fitting results: (a) change trend of the least square error; (b) change trend of the average fitness of the population.

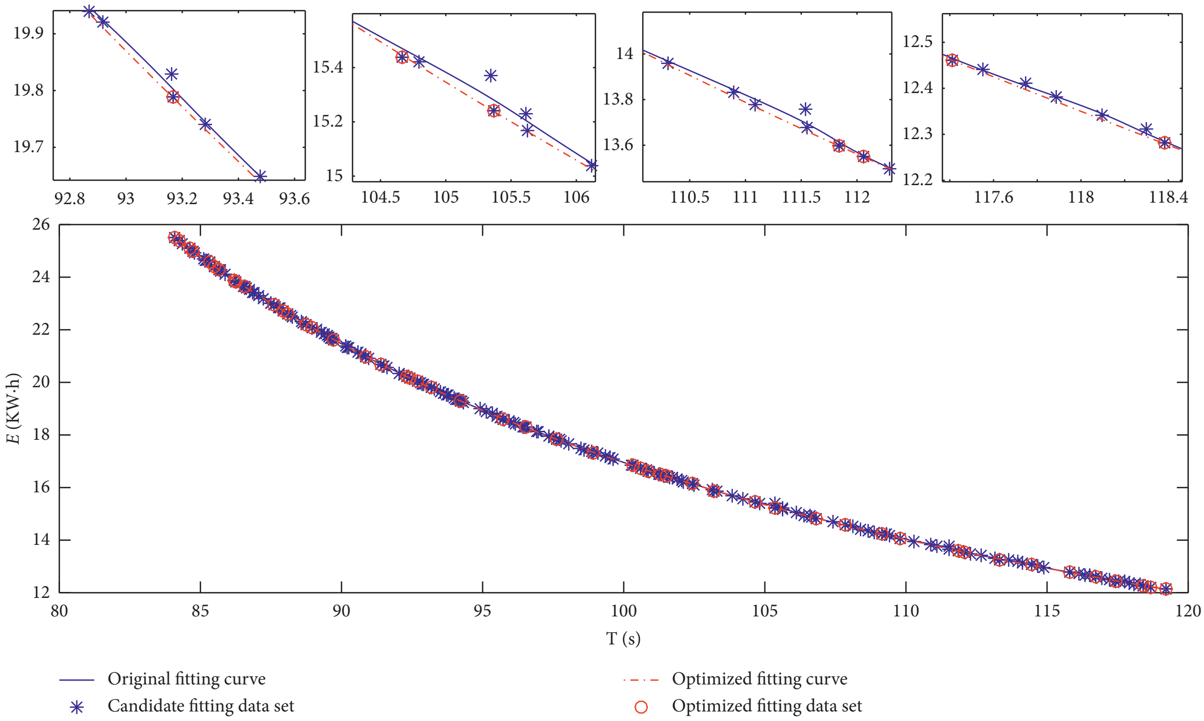


Figure 12: Comparison of fitting curves fitted by different fitting data sets.

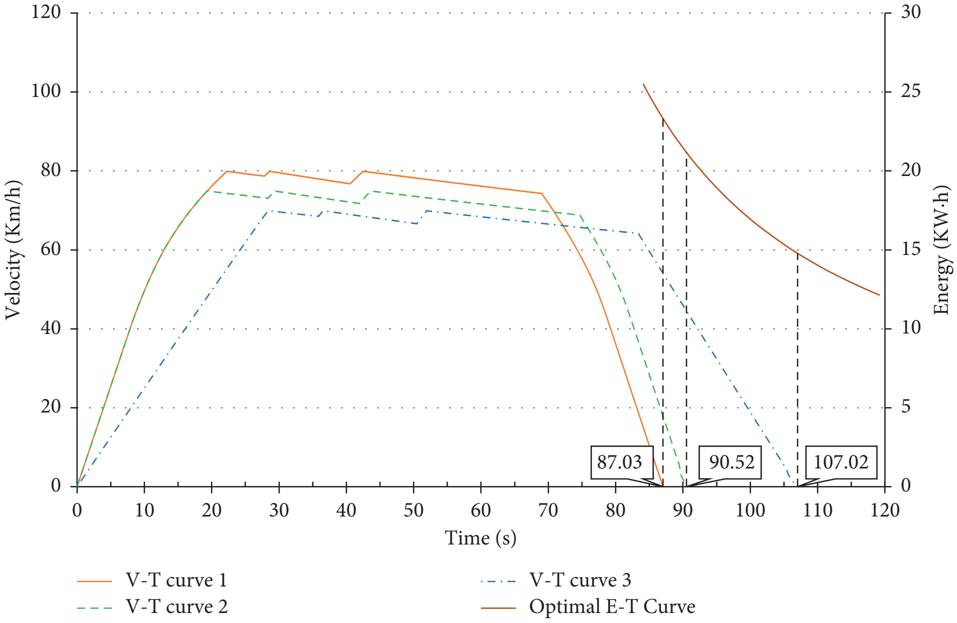


Figure 13: V-T curves corresponding to the optimal E-T points.

randomly select three E-T points from the optimal fitting data set and extract the corresponding V-T curves that have the optimal energy consumption.

The V-S curves with the optimal energy consumption can provide abundant running curve support for the operation of the train under different running time requirements.

6. Conclusions

In this paper, we propose a fitting method of the optimal energy consumption-running time curve of an urban rail section based on train operation data. The main work completed includes the following:

(1) Based on the features of the section operation curve of urban rain trains and correlations between the running time and corresponding energy consumption, the discriminant criterion of outliers is proposed to select the candidate fitting data set from many original data points, which reduces the scale of the train operation data set as well as guaranteeing the curve fitting quality.

(2) An improved B-spline curve fitting method is proposed in which the parameter vector and knot vector are optimized by the genetic algorithm, which has a higher fitting accuracy and faster convergence speed.

(3) On the basis of tabu search, we construct an optimization model of the fitting curve by defining bi-directional tabu lists to adjust and optimize the fitting data set from the candidate data set dynamically. It is proposed that the optimization method could obtain the optimal E-T curve and ensure the continuity and smoothness of the fitting curve at the same time.

The research on the optimal E-T curve based on operation data of urban rain trains has certain practical significance beyond theoretical limitations, and the optimal E-T fitting curve could be used in the selection of section running time, evaluations of the energy consumption of the train operation diagram, and performance appraisal of train drivers. Further research will focus on the optimization ability of the algorithm of fitting curve optimization.

Data Availability

The data used to support the findings of this study 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 Natural Science Foundation Council of China (nos. U1934216 and 71871226) and by the Research Foundation of Guangzhou Metro Group Co., Ltd. (no. HT170235).

References

[1] I. P. Milroy, *Aspects of Automatic Train Control*, Loughborough University, Loughborough, UK, 1980.

[2] E. Khmelnitsky, "On an optimal control problem of train operation," *IEEE Transactions on Automatic Control*, vol. 45, no. 7, pp. 1257–1266, 2000.

[3] C. S. Chang and S. S. Sim, "Optimising train movements through coast control using genetic algorithms," *IEE Proceedings - Electric Power Applications*, vol. 144, no. 1, pp. 65–73, 1997.

[4] B. R. Ke, *Block-layout Design Using MAX-MIN Ant System for Saving Energy on Mass Rapid Transit Systems*, IEEE Press, New York, NY, USA, 2014.

[5] R. Liu and I. M. Golovitcher, "Energy-efficient operation of rail vehicles," *Transportation Research Part A: Policy and Practice*, vol. 37, no. 10, pp. 917–932, 2003.

[6] I. V. Sanchis and P. S. Zuriaga, "An energy-efficient Metro speed profiles for energy savings: application to the valencia Metro," *Transportation Research Procedia*, vol. 18, pp. 226–233, 2016.

[7] Y. V. Bocharnikov, A. M. Tobias, C. Roberts, S. Hillmansen, and C. J. Goodman, "Optimal driving strategy for traction energy saving on DC suburban railways," *IET Electric Power Applications*, vol. 1, no. 5, pp. 675–682, 2007.

[8] M. R. Mashinchi, A. Selamat, S. Ibrahim et al., "Outlier elimination using granular box regression," *Information Fusion*, vol. 27, pp. 161–169, 2016.

[9] H. Hassani, M. Zokaei, D. von Rosen, S. Amiri, and M. Ghodsi, "Does noise reduction matter for curve fitting in growth curve models?" *Computer Methods and Programs in Biomedicine*, vol. 96, no. 3, pp. 173–181, 2009.

[10] S. Zheng, R. Feng, and A. Huang, "A modified moving least-squares suitable for scattered data fitting with outliers," *Journal of Computational and Applied Mathematics*, vol. 370, p. 112655, 2020.

[11] P. Chen, L. Li, Y. Liu et al., "Detection of outliers and patches in bilinear time series models," *Mathematical Problems in Engineering*, vol. 2010, Article ID 580583, 256 pages, 2010.

[12] A. Gálvez and A. Iglesias, "Firefly algorithm for explicit B-spline curve fitting to data points," *Mathematical Problems in Engineering*, vol. 2013, pp. 206–226, 2013.

[13] G. Trejo-Caballero, H. Rostro-Gonzalez, C. H. Garcia-Capulin et al., "Automatic curve fitting based on radial basis functions and a hierarchical genetic algorithm," *Mathematical Problems in Engineering*, vol. 2015, Article ID 731207, 14 pages, 2015.

[14] S. Su, T. Tang, X. Li et al., "Optimization of multitrain operations in a subway system," *IEEE Transactions on Intelligent Transportation Systems*, vol. 15, no. 2, pp. 673–684, 2014.

[15] X. Li, C. F. Chien, L. Li et al., "Energy-constraint operation strategy for high-speed railway," *ICIC International*, vol. 8, no. 10, pp. 6569–6583, 2012.

[16] H. Kang, F. Chen, Y. Li, J. Deng, and Z. Yang, "Knot calculation for spline fitting via sparse optimization," *Computer-Aided Design*, vol. 58, pp. 179–188, 2015.

[17] D. D. Hearn, *Computer Graphics with OpenGL*, Publishing House of Electronics Industry, Beijing, China, 2004.

[18] W. Zheng, P. Bo, Y. Liu, and W. Wang, "Fast B-spline curve fitting by L-BFGS," *Computer Aided Geometric Design*, vol. 29, no. 7, pp. 448–462, 2012.

[19] R. Świta and Z. Suszyński, "Thermal image approximation using B-spline surfaces," *International Journal of Thermophysics*, vol. 39, no. 11, p. 127, 2018.

[20] H. SHOU and L. HU, "B-spline curve fitting algorithm for real-coded GA with normal constraint," *Journal of Zhejiang University of Technology*, vol. 44, no. 4, pp. 466–472, 2016.

[21] L. A. Piegl and W. Tiller, "Least-squares B-spline curve approximation with arbitary end derivatives," *Engineering with Computers*, vol. 16, no. 2, pp. 109–116, 2000.

[22] F. Yoshimoto, T. Harada, and Y. Yoshimoto, "Data fitting with a spline using a real-coded genetic algorithm," *Computer-Aided Design*, vol. 35, no. 8, pp. 751–760, 2003.