Hindawi
Journal of Advanced Transportation
Volume 2023, Article ID 6659916, 12 pages
https://doi.org/10.1155/2023/6659916

Research Article

Short-Term Inbound and Outbound Passenger Flow Prediction for New Metro Stations Based on Clustering and Deep Learning

Zihe Wang,[1] Yongsheng Zhang,[1] Enjian Yao,[1] Yue Wang,[1] Juncheng Li,[2] and Jiantao He[2]

[1]*School of Traffic and Transportation, Beijing Jiaotong University, No. 3 Shangyuancun, Haidian District, Beijing 100044, China*
[2]*Guangzhou Metro Group Co. Ltd, Guangzhou 510380, China*

Correspondence should be addressed to Yongsheng Zhang; yshzh@bjtu.edu.cn

Received 25 November 2022; Revised 9 January 2023; Accepted 20 March 2023; Published 14 August 2023

Academic Editor: Wenxiang Li

The rapid expansion of metro networks, e.g., in many cities of China, continuously introduces the operation of new stations every year. Due to the lack of historical data and complicate variations of short-term passenger flow in the early stage of operation, it is difficult to accurately predict inbound and outbound passenger flows of new metro stations in the short term, which would be the database for train scheduling for new stations before operation, dynamic capacity optimization for new stations under operation, short-term prediction of cycle sharing demands near new stations, and so on. Traditional methods usually failed to exactly reflect the complicate rules or were unusable without the new station's historical data. In order to solve the above problems, this paper proposes a short-term inbound and outbound passenger flow prediction model for new metro stations at the early stage of operation by combining the K-means clustering algorithm, an improved spatiotemporal long short-term memory model (Sp-LSTM), and a real-time feedback error model (mean absolute error, MAE), where passenger flows' spatial-temporal characteristics and land-use relevance are considered. The application in Guangzhou Metro, China, where Line 21 is regarded as a new line, shows that the proposed K-Sp-LSTM model has the best prediction accuracy compared with traditional methods.

1. Introduction

With the national strategy of prioritizing public transportation, public transportation has become the main mode of travel for residents in large cities. Urban rail transit is the preferred mode of public transportation for residents to travel within the city due to its high timeliness, speed, and comfort. Especially in China, there are new lines in operation every year. Since the spatial and temporal distribution patterns of metro passenger flows are becoming more and more complex, accurate passenger flow prediction becomes more and more important for precise operation and management [1]. Especially for new stations, due to the lack of historical data and complicate variations of short-term passenger flow in the early stage of operation, it is difficult to accurately predict inbound and outbound passenger flows of new metro stations in short term, which would be the database for train scheduling for new stations before operation, dynamic capacity optimization for new stations under operation, short-term prediction of cycle sharing demands near new stations, and so on [2]. Therefore, accurate short-term prediction (e.g., in a 15 min time interval) of inbound and outbound passenger flow for new metro stations in the early stage of operation (including implementing the prediction before operation) becomes an essential study.

Three major challenges should be considered in the short-term prediction of inbound and outbound passenger flow for new metro stations in the early stage of operation. The first one is the lack of historical data before operation. Even under operation, the sample size is not large enough for representing the complex passenger flow rules at the early

stage. How to develop a deep learning method for new stations without data or without enough data increases the difficulty. The second one is the complicate spatial-temporal characteristics [3, 4] of inbound and outbound passenger flows in the early stage of operation. Temporally, passenger flow is in a growth stage. Spatially, different stations are latently related since inbound passengers will eventually become outbound passengers. Complicate rules of inbound and outbound passenger flows induce the deep learning method. The third one is how to update inputs step by step with more and more collected data in real time. Traditional methods usually failed to exactly reflect the complicate rules or were unusable without the new station's historical data [5].

Therefore, this paper proposes a deep learning-based short-term inbound and outbound passenger flow prediction model for new metro stations at the early stage of operation by combining the K-means clustering algorithm, an improved spatiotemporal long short-term memory model (Sp-LSTM), and a real-time feedback error model (mean absolute error, MAE). The K-means clustering method is introduced to solve the first challenge, i.e., lack of data, by considering the relationships between passenger flow rules and land use around stations. The Sp-LSTM model is introduced for complicate spatial-temporal characteristics of inbound and outbound passenger flows [6]. The real-time feedback error model provides the inputs updating mechanism. The proposed K-Sp-LSTM-MAE prediction model for short-term inbound and outbound passenger flow forecasting of new metro stations will be deeply discussed in the following parts. This method can realize the short-term passenger flow prediction of the newly built subway station and then provide an important reference for the train operation scheme of the new subway line.

The following sections unfold as follows. Section 2 mainly describes existing studies on passenger flow prediction in recent years. In Section 3, we provide the problem statement. Section 4 elaborates on the K-means, LSTM, and Sp-LSTM. In Section 5, the proposed method is used to predict the passenger flows of new metro stations of Line 21 in Guangzhou Metro, China. The final conclusion is summarized in Section 6.

2. Literature Review

Short-term passenger flow prediction is an important application for artificial intelligence algorithms. Traditional prediction models use analytical equations to analyze the relevant factors affecting the prediction variables and finally fit a prediction function for passenger flow. Fitting the prediction function requires iterative training to find the optimal parameters using a large amount of historical data. Conventional models include the basic time-series model [7], the Kalman filter model [8], the support vector set regression forecasting model [9], and the neural network model [10]. Among these methods, neural networks have demonstrated good results for stable regular passenger flow prediction [11]. However, the passenger flow rule of new metro stations is complex and variable. The conventional neural network models are unable to learn the changing rules of their passenger flows directly and accurately.

The passenger flow of metro stations has complete historical data because of the swipe data of passengers so that there are many methods and more mature methods to predict the passenger flow of metro stations. The metro passenger flow is a time-series problem so that the prediction of passenger flow often introduces deep learning [12, 13]. Liu et al. predicted the future metro passenger flow based only on the historical passenger flow of a single metro station; they used the LSTM as the passenger flow prediction method [14]. This method is effective for predicting metro station passenger flow in a simple metro network. The increasing complexity of metro networks has made the prediction of metro passenger flow more complex, combining more influencing factors. Therefore, various neural networks have been introduced into passenger flow studies in order to combine multidimensional influences. Roos et al. had considered not only the time-series characteristic of passenger flow at a metro station but also the spatial correlation of metro networks based on Bayesian networks. Structure expectation is used to study the spatial relationship of the metro network [16]. Zhang and Ma considered that the inbound and outbound passenger flow of a metro station is proportional to its spatial location and temporal factors so that they calibrated this spatiotemporal relationship with proportional parameters in predicting passenger flow [17].

With the rapid development of advanced positioning technology, other technologies can also be used as an aid for metro passenger flow prediction. Fu et al. innovatively combined cell phone data and metro smart card data [18]. Combining cell phone mobile data which is near metro stations can better assist in predicting metro station passenger flow. In particular, mobile phone data can be more useful for short-term prediction of metro station passenger flow when certain events and activities cause sudden and large changes in metro passenger flow. Yang et al. studied the spatiotemporal characteristics of metro station passenger flow in depth and combined such spatiotemporal characteristics with neural networks to create an improved Sp-LSTM model for short-term inbound and outbound passenger flow prediction [6]. Finally, the accurate prediction results were obtained by using the Beijing Metro Airport Line. Li et al. used search engines such as Baidu and 360 to filter metrics related to various metro trips. This paper incorporates deep learning and LSTM to learn multiple objective functions for aiding the prediction of metro passenger flow [19].

The change of passenger flow in a new station is influenced by various factors, such as line location, fare, and land-use relationships around the station. Therefore, the passenger flow prediction of new metro stations cannot be directly adopted from the traditional passenger flow prediction model. In terms of the research field of short-term passenger flow prediction at the initial operation of new metro stations, there are little methods which can be effective at home and abroad. Guang established a passenger flow prediction model for new stations after new line access based on station accessibility indexes [20]. The method also enables

the evaluation of metro transportation organization plans by means of passenger flow data from the AFC system. Cai et al. used the nonset counting theory to predict the weekday passenger flow between the new stations and the existing stations under the new line access conditions [21]. Cheng et al. used a multiple linear regression model to predict the passenger traffic at the initial stage of the new line opening after analyzing the passenger flow pattern of the existing stations [22]. They also calculated the operating intensity of the rail line at different moments of time. Kepaptsoglou et al. focused on analyzing the passenger demand of the residents around the newly opened line and directly investigated the traffic demand along the line [23]. Based on the survey results, the research team developed a passenger flow prediction model for the new line. Yao et al. proposed a short-term inbound and outbound passenger flow prediction method based on improved k-nearest neighbor nonparametric regression for the new stations [24]. The method considers the mechanism of metro passenger flow generation and analyzes the change rule of station passenger flow at the early stage of operation and its correlation with the land use around the station. Based on the clustering of metro stations, they improved the nonparametric regression algorithm by combining the short-term passenger flow characteristics and proposing the short-term inbound and outbound passenger flow prediction method at the early stage of new stations.

Existing studies have analyzed many factors such as the spatiotemporal characteristics of metro passenger flow and land-use types around metro stations. Some studies also consider the traffic demand around metro stations. However, the current research lacks the ability to combine the historical passenger flow data of the existing metro stations and the land-use relationship around the new stations. The purpose of this paper is to put forward a short-term passenger flow prediction model of new metro station combining these two factors.

The major contributions in this paper are concluded as follows:

(1) A short-term inbound and outbound passenger flow prediction method for new stations at the early stage of operation is proposed by combining clustering and deep learning methods to overcome the difficulties of the lack of historical data and the deep understanding of complicate passenger flow rules in the early operation of the new stations

(2) The spatial-temporal characteristics of inbound and outbound passenger flows as well as a real-time updating mechanism based on real-time data collection are incorporated into the deep learning part of the short-term prediction method for new stations

(3) The proposed method is successfully applied to a new line in Guangzhou Metro network to show the superiority of the proposed method.

3. Problem Statement

3.1. Passenger Flow Characteristics of New Metro Stations. The passenger flow of new metro station will rapidly increase at the initial stage of operation. After a period of operation, the change of passenger flow in the new metro station during weekdays will gradually stabilize, as shown in Figure 1. But its passenger flow volatility is still larger compared with other metro stations, as shown in Figure 2. According to the change of passenger flow data, the new metro station's passenger flow will have sudden changes in a short period of time, while it remains volatile over a longer period of time during weekends. The complicate passenger flow rules for the new stations at the initial stage of operation makes short-term passenger flow prediction very difficult. The deep learning method which is able to represent complicate passenger flow rules is expected.

However, the new station lacked historical passenger flow data when it just operated, which makes the application of the deep learning method, which requires a lot of historical data for training and optimization challengeable. Even for traditional methods, short-term passenger flow prediction is also hard to be implemented without enough historical data. This paper tries to construct a database based on the historical passenger flow of existing metro stations according to the inherent relationship between passenger flow rules and land uses. The database is generated by clustering stations with similar land uses, then the new station will be able to utilize the historical data in the database by pattern matching. By this means, enough data will be available for forecasting the passenger flow of new metro stations in the early stage of operation.

3.2. Short-Term Metro Station Passenger Flow Forecast. The AFC system of the metro can record the number of passengers entering and leaving the station in real time. For simplicity, this paper uses a 15-minute time interval to record and predict the inbound and outbound passenger flow of each station. The study data contain date, period, 15-minute metro station traffic data, and station type. Metro station traffic characteristics have obvious temporal characteristics, so the basic sequence data input to the LSTM is the traffic data divided by period. The inbound and outbound passenger flows of a metro station can be expressed by the following equation:

$$x_c = \left\{x_1^i, x_2^i \cdots x_t^i \cdots\right\}, \tag{1}$$

where x_t^i is the inflow/outflow of the station i in the time interval t and c is the metro station category.

The passenger flow of a metro station is not only related to the time period and the surrounding land use relationship but also related to the spatial location. For example, although two metro stations have the same land-use type, the passenger flow in the same time period may be vastly different.

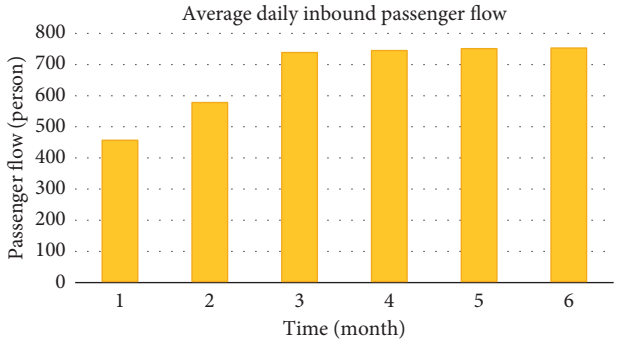

Figure 1: Changes in the average daily entry volume of new metro stations.

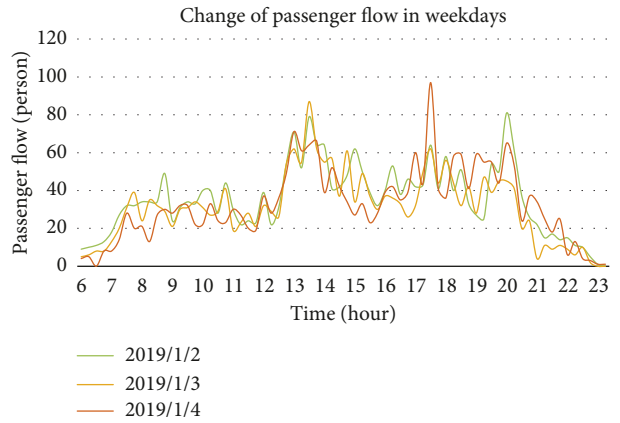

Figure 2: Changes in passenger flow of a new metro station during the initial working days.

Besides the impact of unexpected events, the main reason for the impact is that these two metro stations are located in different locations of the rail network. Therefore, the spatial relationship between stations is also a potential influencing factor for passenger flow forecasting. The spatial influence between metro stations is not measured by simple European distance but is mainly parameterized by the number of interchanges and the train running time between the two stations. Combining the above factors yields a more comprehensive base data matrix X_C, as shown in the following equation:

$$X_C = \begin{bmatrix} x_1^1, x_2^1, \cdots, x_t^1, x_{t+1}^1, \cdots \\ x_1^2, x_2^2, \cdots, x_t^2, x_{t+1}^2, \cdots \\ \cdots\cdots\cdots \\ x_1^i, x_2^i, \cdots, x_t^i, x_{t+1}^i, \cdots \\ \cdots\cdots\cdots \end{bmatrix}. \quad (2)$$

An important feature of new metro stations is that the change in passenger flow is gradually stabilized, i.e., the volatility of passenger flow is reduced over time. It means that it is the short-term data that is important for the update of the model parameters. Therefore, in this paper, MAE is used as a loss function for short-term parameter adjustment. By this means, the change of short-term data could be better captured for the short-term metro station passenger flow prediction, especially when there is an unexpected event in the metro network.

In summary, the short-term passenger flow prediction model for new metro stations needs to take into account both long-term passenger flow characteristics and short-term passenger flow changes as well. In this paper, an improved integrated passenger flow prediction model, K-Sp-LSTM-MAE, is constructed, which takes the land-use relationship around the metro station, the spatial relationship of the metro network, and the historical inbound and outbound passenger flow data to achieve the short-term inbound and outbound passenger flow prediction of the new metro station. The model framework proposed in this paper is shown in Figure 3.

4. Methodology

4.1. Metro Station Clustering Based on K-Means. In order to construct appropriate historical data for the new metro station, this paper proposes to classify stations into different groups with different land uses based on the historical passenger flow data of existing stations. Since morning and evening peak-hour coefficients of inbound and outbound passenger flows reflect land-use property around stations to a great extent, while daily inbound and outbound passenger flow volumes reflect the land-use scale around stations [25], the six variables are utilized as clustering indexes. In this paper, the K-means clustering method algorithm is adopted to cluster metro stations. Since the different types of passenger flow data in this case have vastly different characteristics, K-means clustering yields better results. The K-means clustering algorithm is based on a prototypical, partitioned distance technique, which attempts to discover a user-specified number of clusters.

Based on the surrounding land-use property, in this paper, metro stations are classified as residential, office, office dominant, residential dominant, transportation hub, commercial, and comprehensive use based on morning and evening peak-hour coefficients of inbound and outbound passenger flows. The type of each station may be different between weekdays and weekends. For example, some stations fall into the office occupancy category on weekdays and the commercial category on weekends. Each type of station has a different traffic pattern, as shown in Figure 4, which means that the traffic forecast of a new station needs to be classified according to its land-use relationship first. In the figure, 15-minutes is taken as a time interval, with inbound passenger flow being positive and outbound passenger flow being negative. Based on land-use property clustering results, the daily inbound and outbound passenger flow volumes of a station are then regarded as indexes for the next-step clustering in terms of land-use scale. At last, the stations with the similar land-use property and scale are put into the same group so that the database is generated where land-use property and scale corresponding to historical 15-minutes time-interval-based inbound and outbound passenger flows are recorded. For the above two-step clustering, the K-means method is utilized by this paper.

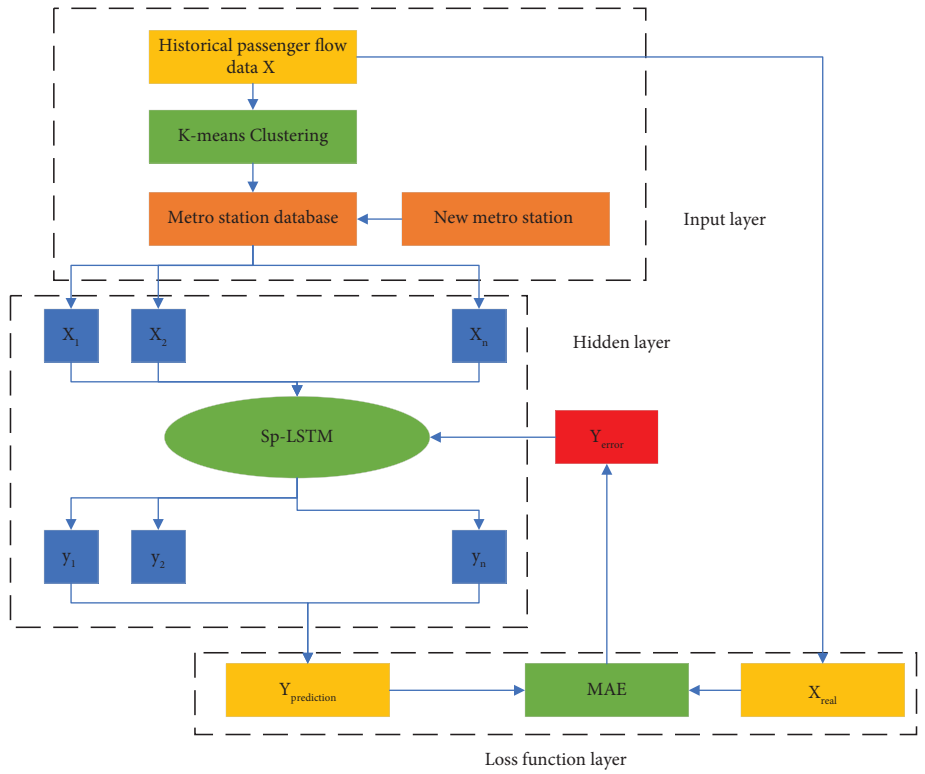

Figure 3: Model framework diagram.

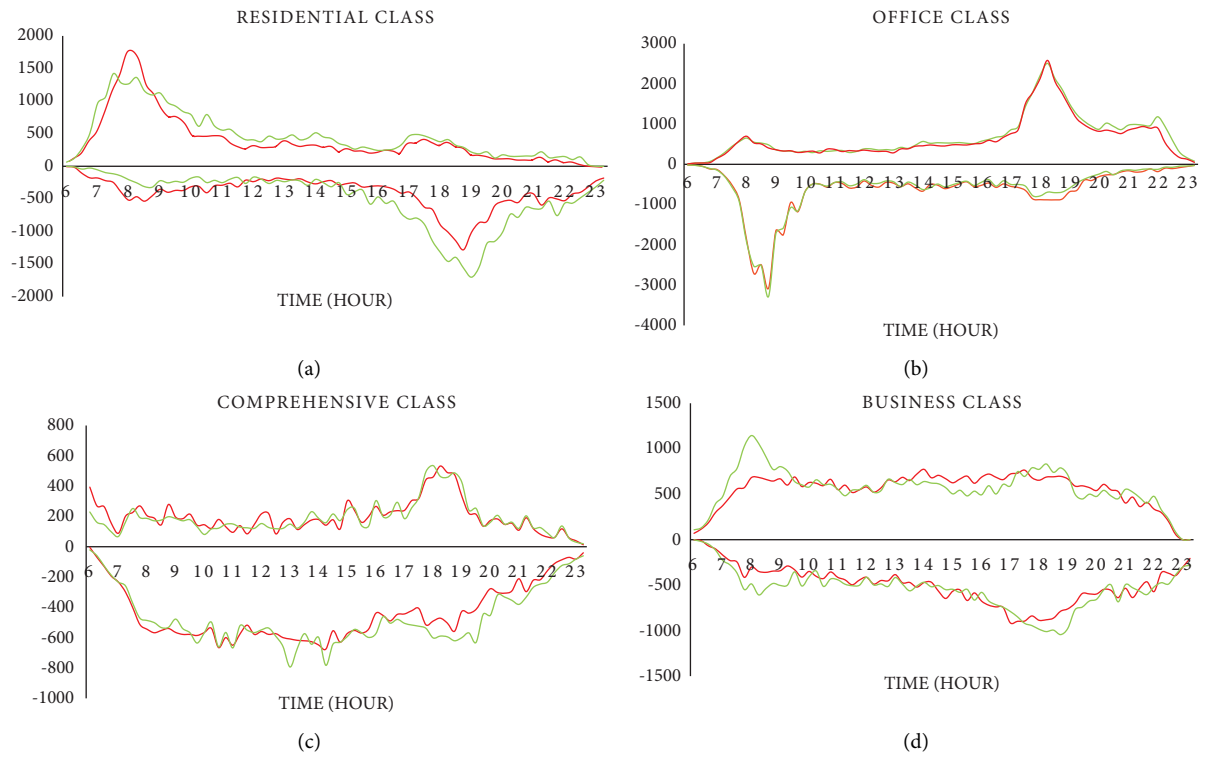

Figure 4: Passenger flow characteristics of different types of metro stations.

In the K-means-based clustering, the greater the distance between two samples, the less similar they are, and the smaller the distance between two samples, the more similar they are. The sample data not only need the land-use type of the station but also need to consider the passenger flow scale of the metro station. Thus, this paper uses Euclidean distance to calculate the two-dimensional spatial distance, as shown in the following equation:

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (y_i - x_i)^2}, \tag{3}$$

where y_i represents the passenger flow of the new metro station during period i. x_i represents the passenger flow of the existing metro station during period i.

First, for the dataset with n data points from n data points, randomly select k points as the initial clustering center and then classify the data near each clustering center by Euclidean distance calculation, using an iterative approach, and continuously update the clustering center points in the above process, until the overall reach a stable state and obtain the clustering center C.

The basic steps of the K-means clustering algorithm calculation are as follows:

Step 1: Input a dataset and determine the number of clustering centers K

Step 2: Generate k initial clustering centers at random in the dataset

Step 3: Calculate the distance between the cluster centroids and other data points and take the minimum value of the distance

Step 4: Calculate the mean values of all objects in the same cluster and count them and update and replace the calculated results with the new cluster centers

Step 5: Repeat the above steps until the results are stable and the cluster centers no longer change, output.

The stable state in the above step requires setting the quantization objective function and then recalculating the center of mass of each cluster based on the result of the function. Considering the data of Euclidean distance, this paper uses the sum of the squared error (SSE) as the objective function of clustering, as in the following equation, with two different sets of clusters generated by running K-means twice. Then, when updating the clustering centers, we choose the cluster center with a smaller SSE.

$$\mathrm{SSE} = \sum_{i=1}^{K} \sum_{x \in C} d(c_i, x)^2, \tag{4}$$

where K means the cluster center, c_i denotes the center of number i, and d is the Euclidean distance function.

4.2. Passenger Flow Prediction Model Combining Temporal and Spatial Characteristics. LSTM has a strong ability to learn the change rule of stable time-series data so that the neural network is used in the study method of historical passenger flow rule in this paper [26]. The gating structure in LSTM can screen out the abrupt passenger flow data very well. This method is suitable for the passenger flow prediction of new metro stations after generating pseudohistorical data for the new stations from the above database by pattern matching.

LSTM adds the state c, called cell state, to the RNN network to preserve the long-term state. As shown in Figure 5, in addition to h flowing with time, cell state c also flows with time during the process, and cell state c represents long-term memory.

The internal structure of the LSTM is shown in Figure 6, where the forgetting gate determines how much of the cell state c_{t-1} of the previous moment is kept to the current moment c_t; the input gate determines how much of the input x_t of the network at the current moment is saved to the cell state c_t; and the output gate to control how much of the cell state c_t is output to the current output value h_t of the LSTM.

First, the network input h_{t-1} at moment $t-1$ is combined with the network input x_t at this step, and after doing a linear transformation, the result is mapped to 0~1 as the decay coefficient of memory after a sigmoid activation function of σ, noted as f_t.

$$f_t = \sigma\left(W_f \left[h_{t-1}, x_t\right] + b_f\right). \tag{5}$$

The sigmoid function, called the input gate, determines what values are to be updated and the tanh layer creates a new vector of candidate values $\widetilde{c_i}$ to be added to the state.

$$\begin{aligned} i_t &= \sigma\left(W_i \left[h_{t-1}, x_t\right] + b_i\right), \\ \widetilde{c_i} &= \tanh\left(W_c \left[h_{t-1}, x_t\right] + b_c\right). \end{aligned} \tag{6}$$

The cell state c_t at the current moment is obtained by multiplying the previous cell state c_{t-1} by the original element by the forgetting gate f_t and then multiplying the current input cell state $\widetilde{c_i}$ by the input gate i_t and then adding these two factors together.

$$c_t = f_t * c_{t-1} + i_t * \widetilde{c_i}. \tag{7}$$

The output gate controls the effect of long-term memory on the current output, as determined by both the output gate and the cell state.

$$\begin{aligned} o_t &= \tanh\left(W_o \left[h_{t-1}, x_t\right] + b_o\right), \\ h_t &= o_t * \tanh\left(c_t\right). \end{aligned} \tag{8}$$

LSTM is very capable of learning one-dimensional serial data. In the field of rail passenger flow prediction, LSTM can accurately learn the intrinsic connection between the input passenger flows for each time period. It is easy to find that LSTM learns and predicts the passenger flow characteristics of stations. However, the inbound and outbound passenger flows of metro stations in a rail network are closely connected. The interaction between metro stations is important for passenger flow prediction. Especially when unexpected events cause unconventional changes in passenger flow, relying on the spatial influence relationship of metro stations for short-term passenger flow prediction is more sensitive

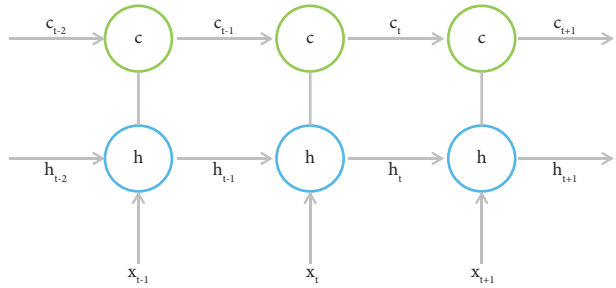

Figure 5: LSTM network structure.

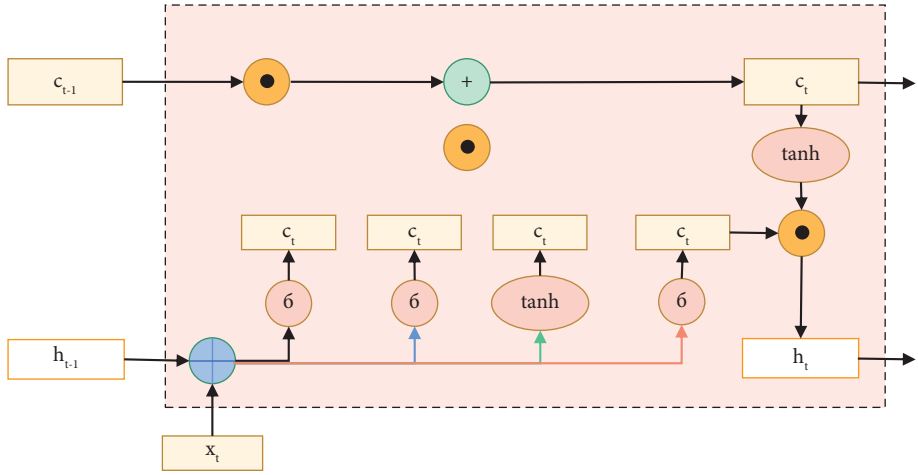

Figure 6: LSTM internal structure.

than conventional time-series prediction. Therefore, the lack of multidimensional input of LSTM leads to its inability to take into account multiple factors. Yang et al. considered the spatial influence between metro stations and proposed an improved Sp-LSTM [6]. In this paper, the parameters are adjusted according to the passenger flow characteristics of new metro stations.

Spatiotemporal long- and short-term memory network (Sp-LSTM) takes into account the time delay from inbound to outbound passenger flow. It is able to capture the spatial influence of inbound volume on outbound volume. On this basis, multidimensional sequences are established together as feature sequences to achieve short-term intelligent prediction of outbound station volume.

In metro passenger flow prediction, each passenger flow index is regarded as a sequence of data that changes with time. It has a certain relationship with its past passenger flow. By analyzing the passenger flow data collected in real time and using the regularity of passenger flow itself to mine and extract the features of passenger flow, the short-term intelligent prediction of passenger flow can be realized.

We set two parameters m and n to measure the relationship between the inbound and outbound passenger flow of two metro stations. We assume that the passenger flow between the target metro station k and other metro station i is $f_{i,k}$, and the outbound passenger flow of the target metro stations k is f_k. m_i means the outbound ridership contribution of station i to station k, as shown in the following equation:

$$m_i = \frac{f_{i,k}}{f_k}. \tag{9}$$

We assume that the inbound passenger flow of target metro stations k is $f_k^{'}$. n_i means the inbound ridership contribution of station i to station k, as shown in the following equation:

$$n_i = \frac{f_{i,k}}{f_k^{'}}, \tag{10}$$

$$u_i = \frac{m_{\max} - m_i}{m_{\max} - m_{\min}}, \tag{11}$$

$$v_i = \frac{n_{\max} - n_i}{n_{\max} - n_{\min}}, \tag{12}$$

$$Z_k = w_1 \otimes U + w_2 \otimes V, \tag{13}$$

where $\otimes$ is Hadamard product, Z_k is the influence coefficient matrix of metro station k, and other metro stations.

4.3. Short-Term Updating Mechanism. In this paper, the K-Sp-LSTM model is used to forecast the short-term inbound and outbound passenger flow of new metro stations. Newly

built metro stations have high volatility of passenger flow so that the short-term passenger flow data of the day is more valuable. This paper finally adopts MSE as the loss function to learn the short-term change characteristics of short-term passenger flow, as shown in equation (13). MAE can better reflect the absolute value of change, which is more suitable for passenger flow change analysis. It corresponds to that the change of MAPE and other indicators is too large when the passenger flow in metro stations is relatively small. The short-term passenger flow and prediction results are input into the loss function to iteratively train the weight parameter w and bias parameter b. The error results will be returned to the deep learning model for short-term passenger flow change feature learning.

$$L = \frac{1}{n} \times \sum_{i=1}^{n} \left(w\widehat{y}_i + b - y_i \right)^2, \tag{14}$$

where L is loss function; w represents the weight parameter; and b is the Offset parameter. y_i is the real passenger flow of objected station i; $\widehat{y}_i$ is the predicted passenger flow of objected station i; n is the total number of new stations; and i is a particular station.

5. Case Study

5.1. Data Analysis. The data selected in this paper are the smart card data recorded by the AFC system in all metro stations in the Guangzhou rail network. Passenger flow time period is from 6 : 00 to 23 : 30 with a time interval of 15 minutes. The new metro stations selected for the study are the eight metro stations of Guangzhou Metro Line 21 (Zhenlongxi Station to Zengchengguangchang Station). Guangzhou Metro Line 21 is the 14th line built and operated by the Guangzhou Metro, which opened for operation on December 28, 2018, for the first section of the project. The first section of the line has only one metro interchange station. The line is located in the northeastern part of Guangzhou City, as shown in Figure 7, with an "L" shape. The line is 62.6 kilometers long and covers a number of functional areas, including residential, commercial, and industrial areas. It promotes the development of the eastern part of the city and the surrounding suburbs.

The time frame studied in this paper is the daily inbound and outbound passenger flow of all metro stations from November 1, 2018, to April 30, 2019. Among them, the metro stations of the first section of Guangzhou Metro Line 21 have smart card data starting from December 28, 2018.

5.2. Clustering Results. This study uses the K-means clustering method to cluster all metro stations in Guangzhou, as shown in Table 1. The indexes of clustering are the peak-hour coefficient and the total number of inbound and outbound stations. The peak-hour coefficient can reflect the land-use relationship around the metro station. The number of passengers entering and leaving the station throughout the day reflects the scale of passenger flow in the metro

Figure 7: Guangzhou Metro network.

station. The morning and evening peak hours are 7 : 00–9 : 00 and 17 : 00–19 : 00, respectively. The ratio between the passenger flow and the daily average total passenger flow in the two periods is regarded as the morning and evening peak coefficient. Based on the land-use relationship around the newly built metro stations, this paper classifies these stations as shown in Table 2. In this paper, the correlation analysis of passenger flow between existing and newly built metro stations of various types is shown in Table 3. Through several common correlation tests, it can conclude that the clustering effect is good.

As can be seen from the correlation coefficient results in the table, this study accurately matched the types of new metro stations and existing metro stations. In this study, the passenger flow of newly built metro stations and the passenger flow of various types of metro stations in the historical database were compared by MAPE, as shown in Table 4. Meanwhile, this paper compares the passenger flow of a newly built metro station with that of various metro stations, as shown in Figure 8. The red lines represent the inbound and outbound passenger flow of existing residential stations; the green lines represent the inbound and outbound passenger flow of the new residential stations; and the yellow lines represent the inbound and outbound passenger flow of residential stations.

5.3. Analysis of Model Prediction Accuracy. In this paper, the ridership of five metro stations of Guangzhou Metro Line 21 is used in the training set, and the ridership of the remaining three metro stations is regarded as the test set.

This paper mainly selects the following three evaluation indicators for error analysis: mean absolute percentage error (MAPE), mean absolute error (MAE), and weighted mean absolute percentage error (WMAPE). The equations are as follows:

Table 1: Clustering results of metro stations.

Type of metro station	Metro station
Office class	Xinnan station, Zhenlong station, Tiyuzhongxin station...
Residential class	Fengxia station, Zhishicheng station, Sanyuanli station...
...	...
Comprehensive class	Zhenlongbei station, Wushan station, Jiangnanxi station...
Commercial class	Wangcun station, Tianhegongyuan station...

Table 2: Clustering results of new metro stations.

Type of metro station	New station
Office class	Kengbei station
Residential class	Shantian station, Zhongxin station, Fenggang station
Comprehensive class	Zhenlongxi station
Commercial class	Zengchengguangchang station, Zhucun station

Table 3: Correlation analysis of passenger flow between new and existing stations.

Correlation coefficient	Residential class	Office class	Business class	Comprehensive class
Pearson's	0.822	0.817	0.818	0.782
Kendall's	0.581	0.564	0.560	0.536
Spearman's	0.727	0.714	0.715	0.703

Table 4: MAPE between the new stations and the existing metro stations.

Station	Residential class (%)	Office class (%)	Business class (%)	Comprehensive class (%)
Zhenlongxi station	28.32	24.08	25.17	**15.46**
Zhongxin station	**13.21**	28.41	29.31	23.74
Zhucun station	31.42	29.04	**12.82**	26.95
Kengbei station	**9.94**	24.32	25.71	21.72

The bold values show the importance of cluster prediction for different types of metro stations. The accuracy of passenger flow prediction based on the same type of metro station is obviously higher than other types.

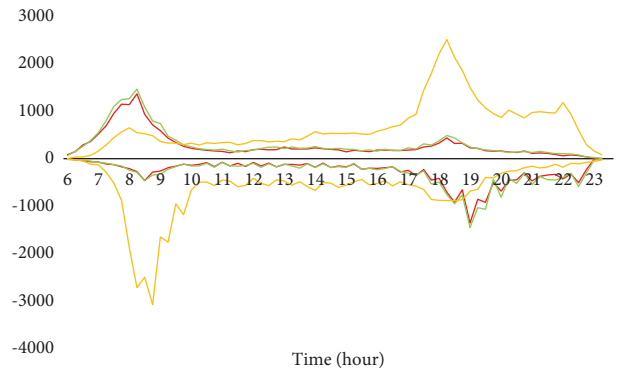

Figure 8: Passenger flow comparison between the new station and the two types of stations.

Table 5: Model parameter values and error results.

Epochs	Hidden units	MAE	MAPE (%)
500	100	11.13	8.32
	500	11.04	8.29
1000	100	10.97	8.27
	500	10.54	8.25

Table 6: Comparison of error analysis with traditional models.

Model	MAPE (%)	MAE	WMAPE (%)
CNN	12.18	15.62	6.21
LSTM	17.97	20.13	7.94
Sp-LSTM	8.31	11.32	4.78
K-Sp-LSTM-MAE	8.25	10.54	4.29

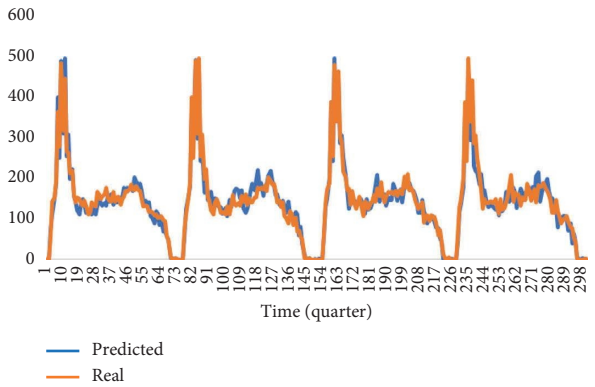

Figure 9: Comparison of predicted and true values for working days.

$$\begin{aligned}
\text{MAPE} &= \frac{100}{n} \times \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right|, \\
\text{MAE} &= \frac{1}{n} \times \sum_{i=1}^{n} |y_i - \hat{y}_i|, \\
\text{WMAPE} &= \sum_{i=1}^{n} \left(\frac{y_i}{\sum_{j=1}^{n} y_j} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \right), \\
\text{MSE} &= \frac{1}{n} \times \sum_{i=1}^{n} (\hat{y}_i - y_i)^2,
\end{aligned} \tag{15}$$

where MAPE is the average absolute percentage error; MAE is the mean absolute error; WMAPE is the weighted average absolute percentage error; MSE is the mean squared error. y_i is the real passenger flow of station i; $\hat{y}_i$ is the predicted passenger flow of station i; n is the total number of stations; and i is a particular station.

For the K-Sp-LSTM-MAE model, we need to adjust the two parameters, epoch and hidden unit, to make the best prediction of the model. The results of the parameter adjustment are shown in Table 5. From the table, we can see that the model achieves the best prediction when Epoch- = 1000 and hidden units = 500. In addition, other prediction models such as CNN, LSTM, and Sp-LSTM are compared in this paper, and finally, K-Sp-LSTM-MAE has more accurate results, as shown in Table 6.

According to the clustering prediction results, the passenger flow prediction on weekdays is better than that on weekends, and this paper predicts the passenger flow of metro stations on weekdays and weekends, as shown in Figures 9 and 10, respectively. The main function of the

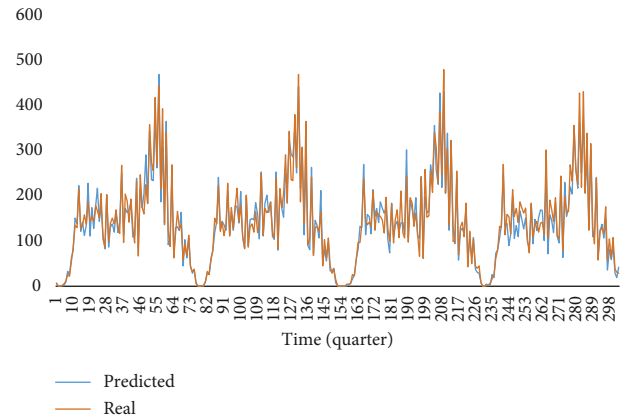

Figure 10: Comparison of predicted and true values for weekends.

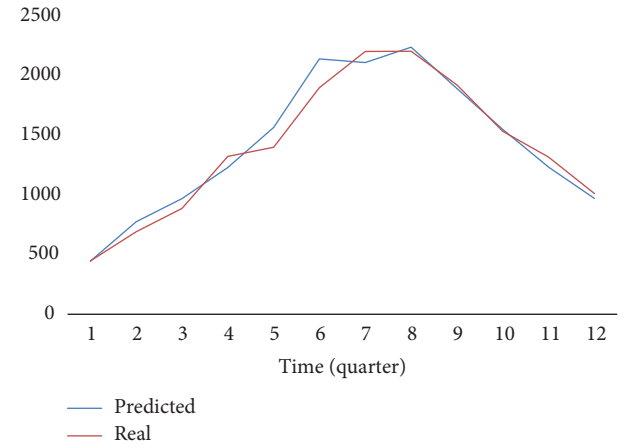

Figure 11: Comparison between predicted and real values of morning peak on working days.

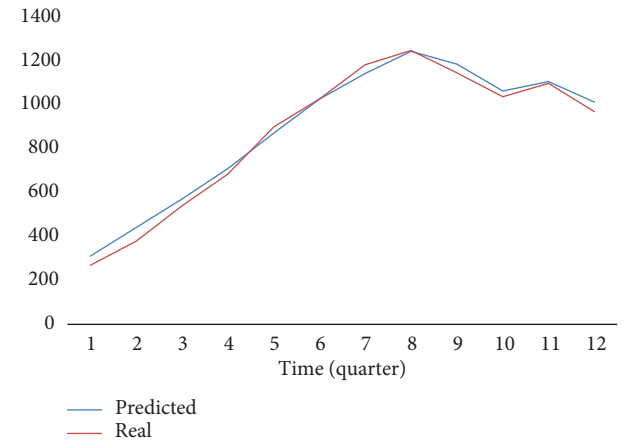

Figure 12: Weekend morning peak forecast versus true value.

passenger flow prediction of metro stations is to plan the service resources of metro stations in advance and provide sufficient capacity resources. Therefore, this paper focuses on learning and predicting the passenger flow pattern during morning peak hours, as shown in Figures 11 and 12.

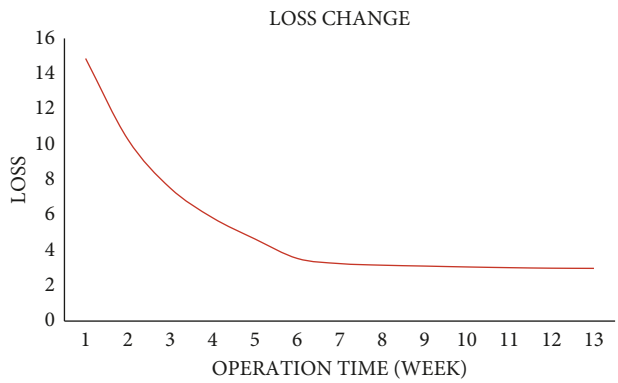

Figure 13: Variation of loss function.

From the above results, it can be concluded that the K-Sp-LSTM-MAE model can accurately classify new metro stations and predict passenger flow according to the type of metro stations. The new metro stations will be in the growth stage of passenger flow in the initial period of opening. Therefore, the value of the loss function in this model will also be larger in the four weeks before opening, as shown in Figure 13. As the operation time of Guangzhou Metro Line 21 increases, the passenger flow pattern of each metro station will gradually stabilize, and the prediction accuracy of the model will also be improved.

6. Conclusion

The purpose of this paper is to predict the short-term inbound and outbound passenger flows of newly built metro stations at the initial stage of operation. This paper combines the spatial relationship of each metro station in the metro network, the temporal properties of metro passenger flow, and the land-use relationship around the metro stations. This paper innovatively proposes a K-Sp-LSTM-MAE model. The model takes the historical inbound and outbound passenger flow of metro stations as the main input and analyzes the spatiotemporal properties of passenger flow in new metro stations. Finally, we validate the model with Guangzhou Metro Line 21 as an example. The results show that the K-Sp-LSTM-MAE model outperforms the traditional CNN, LSTM, and Sp-LSTM in prediction.

The short-term passenger flow prediction method proposed in this paper is not only applicable to the passenger flow in and out of new metro stations but also the new metro lines. This method combines clustering analysis, neural network, and metro passenger flow characteristics. At the same time, the short-term adjustment mechanism is also introduced in this paper so that the model can still maintain high prediction accuracy when the metro passenger flow changes unconventionally. The passenger flow prediction of new metro stations can provide an important basis for the resource allocation of metro stations and the formulation of train diagrams.

The prediction of short-term inbound and outbound passenger flow at the early stage of operation of new metro stations is important for the organization and management of metro stations, the preparation of train operation diagrams, and the number of shared bikes placed around metro stations. Due to the lack of historical passenger flow data in new metro stations and the unstable rules of passenger flow changes in the early stage of operation, more factors need to be taken into account in the short-term inbound and outbound passenger flow prediction. In this paper, we use the clustering method to learn the passenger flow patterns of metro stations of different land-use types on weekdays and weekends and then use a neural network considering the spatial relationship between metro stations for preliminary passenger flow prediction. Considering the volatility of passenger flow in new metro stations, an instant passenger flow error adjustment mechanism is also introduced in this paper. Finally, this paper establishes a complete short-term passenger flow prediction system for new metro stations.

The model proposed in this paper still has several shortcomings. The model lacks the investigation of passengers' travel demand. The relationship between the inbound and outbound passenger flow of metro stations and the OD passenger flow between metro stations are not fully utilized in this paper. In the future, the survey data of passenger travel choice should be introduced into the short-term passenger flow prediction of new metro stations, which can help the prediction model to further improve the prediction accuracy.

Data Availability

Some or 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 that there are no conflicts of interest regarding the publication of this article.

Authors' Contributions

Zihe Wang, Yongsheng Zhang, and Enjian Yao conceptualized and designed the study. Juncheng Li and Jiantao He collected the data. Yongsheng Zhang and Yue Wang performed analysis and interpreted the results. Zihe Wang and Yongsheng Zhang prepared the draft. Juncheng Li and Jiantao He modified the draft. All authors have reviewed the results and approved the final version of the manuscript.

Acknowledgments

This research was supported by the National Natural Science Foundation of China (nos. 52102387 and 52172312).

References

[1] T. Lyu, M. Xu, J. Zhang, Y. Wang, L. Yang, and Y. Gao, "Influential factor analysis and prediction on initial metro network ridership in xi'an, China," *Journal of Advanced Transportation*, vol. 2022, Article ID 2842949, 18 pages, 2022.

[2] W. Li, S. Chen, J. Dong, and J. Wu, "Exploring the spatial variations of transfer distances between dockless bike-sharing systems and metros," *Journal of Transport Geography*, vol. 92, Article ID 103032, 2021.

[3] C. Xie, X. Li, B. Chen, F. Lin, Y. Lin, and H. Huang, "Subway sudden passenger flow prediction method based on two factors: case study of the dongsishitiao station in beijing," *Journal of Advanced Transportation*, vol. 2021, Article ID 2842949, 8 pages, 2021.

[4] Q. Tang, M. Yang, and Y. Yang, "ST-LSTM: a deep learning approach combined spatio-temporal features for short-term forecast in rail transit," *Journal of Advanced Transportation*, vol. 2019, Article ID 2842949, 8 pages, 2019.

[5] T. Lu, E. Yao, S. Liu, and W. Zhou, "Short-term Prediction of passenger flow in and out of new urban rail lines," *Journal of the Chinese Railway Society*, no. 05, pp. 19–28, 2020.

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

[7] J. G. De Gooijer, R. J. Hyndman, and J. Rob, "25 Years of time series forecasting," *International Journal of Forecasting*, vol. 22, no. 3, pp. 443–473, 2006.

[8] I. Okutani and Y. J. Stephanedes, "Dynamic prediction of traffic volume through kalman filtering theory," *Transportation Research Part B: Methodological*, vol. 18, no. 1, pp. 1–11, 1984.

[9] Y. Sun, B. Leng, and W. Guan, "A novel wavelet-SVM short-time passenger flow prediction in Beijing subway system," *Neurocomputing*, vol. 166, pp. 109–121, 2015.

[10] W. Yu and C. Mu, "Forecasting the short-term metro passenger flow with empirical mode decomposition," *Transportation Research Part C*, vol. 22, pp. 148–162, 2012.

[11] S. Dong, *The Research of Short–Time Passenger Flow Forecasting Based on Improved BP Neural Network in Urban Rail Transit*, Beijing Jiaotong University, Beijing, China, 2013.

[12] N. Huang, Z. Shen, S. Long et al., "The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. In: proceedings of the Royal Society of London," *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, vol. 454, no. 1971, pp. 903–995, 1998.

[13] J. Zhang, Y. Zheng, and D. Qi, "Deep spatio-temporal residual networks for citywide crowd flows prediction," in *Proceedings of the Conference on Artificial Intelligence*, pp. 1655–1661, Hong Kong, China, November 2016.

[14] Y. Liu, Z. Liu, and R. Jia, "DeepPF: a deep learning based architecture for metro passenger flow prediction," *Transportation Research Part C: Emerging Technologies*, vol. 101, pp. 18–34, 2019.

[15] N. G. Polson and V. O. Sokolov, "Deep learning for short-term traffic flow prediction," *Transportation Research Part C: Emerging Technologies*, vol. 79, pp. 1–17, 2017.

[16] J. Roos, G. Gavin, and S. Bonnevay, "A dynamic Bayesian network approach to forecast short-term urban rail passenger flows with incomplete data," *Transportation Research Procedia*, vol. 26, pp. 53–61, 2017.

[17] H. Zhang and W. Ma, "Metro passenger flow forecasting model based on temporal and spatial characteristics," *Computer science*, vol. 46, no. 7, pp. 292–299, 2019.

[18] X. Fu, Y. Zuo, J. Wu, Y. Yuan, and S. Wang, "Short-term prediction of metro passenger flow with multi-source data: a neural network model fusing spatial and temporal features," *Tunnelling and Underground Space Technology*, vol. 124, Article ID 104486, 2022.

[19] H. Li, K. Jin, S. Sun, X. Jia, and Y. Li, "Metro passenger flow forecasting though multi-source time-series fusion: an ensemble deep learning approach," *Applied Soft Computing*, vol. 120, Article ID 108644, 2022.

[20] Z. Guang, *Urban Rail Transit Passenger Flow Prediction Based on Land Use and Accessibility*, Beijing Jiaotong University, Beijing China, 2013.

[21] C. Cai, E. Yao, and Y. Zhang, "Prediction of passenger flow distribution between urban rail stations based on AFC data," *China Railway Science*, vol. 36, pp. 126–132, 2015.

[22] T. Cheng, F. Zhou, and H. Li, "Passenger flow prediction of the southern section of xi 'an metro line 2 in the initial operation," *Urban Rapid Rail Transit*, vol. 28, no. 5, pp. 45–49, 2015.

[23] K. Kepaptsoglou, A. Stathopoulos, and M. G. Karlaftis, "Ridership estimation of a new LRT system: direct demand model approach," *Journal of Transport Geography*, vol. 58, pp. 146–156, 2017.

[24] E. Yao, W. Zhou, and Y. Zhang, "Short-term Passenger flow prediction of new urban rail transit station in the early opening period," *China Railway Science*, vol. 2, pp. 119–127, 2018.

[25] M. Tu, W. Li, O. Orfila, Y. Li, and D. Gruyer, "Exploring nonlinear effects of the built environment on ridesplitting: evidence from Chengdu," *Transportation Research Part D: Transport and Environment*, vol. 93, Article ID 102776, 2021.

[26] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.