Proceedings of the 2nd International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/102/20241020

Short-term passenger flow prediction of urban rail transit based on ARIMA model

Shangyun Li[1,3,*], Ziyi Wang[2]

[1]Suzhou Industrial Park Xinghai Experimental Middle School, Suzhou, 215000, China
[2]Beijing No. 35 Middle School, Beijing, 100000, China

[3]Lsy022901@outlook.com
*corresponding author

Abstract. This paper is about the short-term passenger flow prediction of urban rail transit based on the ARIMA model. Due to the acceleration of urbanization and the increasing density of urban rail transit, more and more people are taking urban rail transit. However, excessive short-time passenger flow can lead to traffic accidents. Therefore, it is very important to predict short-term passenger flow. This paper aims to use the ARIMA model to predict the passenger flow in a short period of time in urban rail transit. A competition data set from Alibaba about the short-time passenger flow of a metro station in Shenzhen was adopted and analyzed by the ARIMA model. The data set contains 1,428 samples, and the time granularity is 10 minutes. After research, it has been found that the ARIMA model can basically predict the short-time passenger flow of the metro with accuracy under the appropriate parameter setting. This also means that the application of the ARIMA model can be extended to other aspects of short-time passenger flow prediction, not just the metro.

Keywords: ARIMA model, short-term passenger flow, prediction.

1. Introduction

With the growth of urban population and the acceleration of urbanization, the passenger flow carried by the rail transit system is gradually increasing. Meanwhile the whole rail transit system is suffering from floods of passengers. By the end of 2022, the length of rail transit lines in China had exceeded 10 000 km [1]. While the efficient and convenient rail transit network brings convenience to residents, it also produces problems such as increased congestion and reduced travel quality [2]. Such problems are putting passengers in anxiety of being late and even causing some serious traffic accidents indeed. A lot of scholars are making efforts to reduce such anxiety and accidents by conducting researches and offering practice solutions. As a consequence, it is of importance to improve transportation efficiency, reduce congestion and delays, and improve passenger travel experience. Of equal importance are analyzing the influencing factors of various types of inbound passenger flow at relevant stations and accurately grasping the distribution characteristics of passenger flow under different travel needs. However, it will be useless to singly analyze the history statistics but not predict the future trends if improvements on the experience of urban rail transit are in urgent need. For this reason, data about urban rail transit are collected so that this paper can base on them. At the same time, for the passenger flow that is about to enter the station, it is necessary to construct a prediction model to accurately predict

DOI: 10.54254/2755-2721/102/20241020

which can provide more powerful decision support for the operation management of urban rail transit, emergency dispatching, and has important theoretical significance and practical value [2].

Currently, there are 3 commonly used models for short-term passenger flow predicting [3]: Traditional time series models, such as moving averages (MA), Auto-regressive Integrated Moving Average (ARIMA), Kalman filtering (KF). Machine learning models, such as support vector machines (SVM), Recurrent Neural Network (RNN), Long short-term memory networks (LSTM), Convolution Neural Network (CNN) and so on [4].

The research process of urban traffic passenger flow prediction can be simplified into three stages: statistical methods, traditional machine learning algorithms, and deep learning algorithms. In the statistical method stage, researchers pay more attention to capturing the linear relationship between variables, however, this method cannot fully grasp the nonlinear relationship in the data. Such methods mainly include Kalman filter model. In the transition to traditional machine learning algorithms, researchers are looking for more comprehensive ways to analyze data and try to extract more valuable information from massive amounts of data. Although these methods have improved in dealing with linear relations, they still have limitations in grasping complex nonlinear relationships [5]. With the rise of deep learning algorithms, researchers have begun to use more advanced technical means to predict passenger flow. Deep learning algorithms are able to more accurately capture nonlinear relationships in the data, so as to more accurately predict the flow of urban traffic passengers. This approach has significant advantages in dealing with complex data relationships and provides strong support for urban transport planning and management [6].

In summary, the three different stages of the method have their own advantages and disadvantages, and together constitute the research process of urban traffic passenger flow prediction. Each approach plays an important role in its field of application and provides a valuable reference for urban traffic management and planning. Auto regressive Integrated Moving Average Model (ARIMA), Logistic Regression (LR), and Grey Model (GM) [7].

Wang proposed a deep learning method for a constitutional long short-term memory network (ConvLSTM) that can consider both multi-level temporal features and spatial features [8]. Zhang et al. proposed a combined model combining Residual Network (Res Net), Graph Constitutional Network (GCN) and long short-term memory network (LME) for prediction, considering the topological relationship, weather conditions, and air quality between metro stations [9]. Zhang proposed a short-term passenger flow prediction method for urban rail transit based on the LSTM model [10]. The model also provides an opportunity to learn the long-term and short-term dependence of passenger flow and external influencing factors and realize multi-scenario prediction such as regular days, holidays, and different weather [10]. In view of such conditions, this paper will contain a method called Auto regressive Integrated Moving Average model (ARIMA).

2. Methodology

2.1. Background introduction

Urban rail transit is a safe, efficient and green emerging public transport mode. It has largely had an effective impact on the contradiction between transportation supply and demand in large and medium-sized cities. The short-term passenger flow of urban rail transit is characterized by poor regularity and significant nonlinear characteristics. In this regard, domestic and foreign scholars' research on short-term passenger flow predicting methods is mainly divided into three categories, namely: parametric model, non-parametric model, combined model [11]. When predicting short-term passenger flow, many methods and models are applied to accurately predict short-term passenger flow of urban rail transit to ensure operation safety and improve transport efficiency. However, some models such as traditional time series models and econometric models have some limitations in time series predicting. For example, traditional models have challenges in handling large-scale data and may face issues such as computational complexity and storage requirements. In addition, the complex nonlinear relationship between the data also makes it difficult to fit the model. Recurrent neural network (RNN) and

DOI: 10.54254/2755-2721/102/20241020

convolution Neural network (CNN) are two commonly used time series prediction methods based on artificial intelligence. RNN is a recursive model that models continuous time points based on Markov assumptions, captures context information and shares parameters. However, this approach also faces some problems, such as gradients that are too large or too small to effectively capture long-term dependencies in the data. This has also led to the emergence and application of its variant Long Short-term memory network (LSTM) [12].

2.2. Data source and prepossessing

First, by adjusting the difference order of the model, the stability of the stochastic process is tested. Then the ARIMA model parameters are estimated. In the research process, when the order is 2, it cannot correspond to the objective data. However, when the order is adjusted to a higher value for operation, the requirements of the white noise test was met. Figure 1 shows the maltreatment of short-time passenger flow in Shenzhen metro stations. It is worth mentioning that the data set used in this paper is a competition data set provided by Ali baba. The data set contains 1428 items, and its time granularity is 10 minutes. It can be easily seen from Figure 1 that the passenger flows vary from day to day. Obviously, it always came as a tip on Mondays which were the beginning of the weekdays and a bottom on Sundays which were the end of the weekends. Additionally, passenger flow data changes regularly on a weekly basis. This means that this set of data is relatively easy to predict (figure 1).

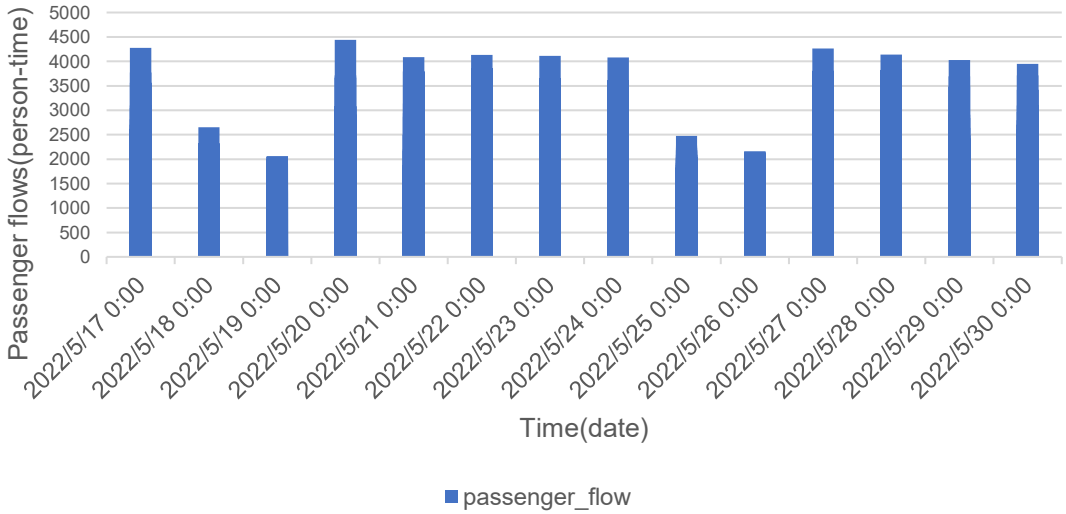

Figure 1. Passenger flow-Time

The first half in data is generally planned and processed, while the other half of the data is used to test the prediction. Soon the short-time passenger flow prediction is carried out.

2.3. Method introduction

However, nowadays the ARIMA prediction model is more accurate in predicting the short-term passenger flow of urban rail transit and can provide more accurate prediction results. In other words, through the auto-regressive difference integrated average moving model (ARIMA), the lag value of the dependent variable, the lag value generating random errors and the current value can be predicted. This method can effectively avoid the problem of the regularity of passenger flow data in short-term passenger flow prediction, and the passenger flow will change greatly with the error. ARIMA model is a famous time series prediction method, which was proposed by Box and Jenkins in the 1970s. The opinion of Box and Jenkins showed that the current value of a time series can be explained by its own historical value as well as the random disturbance term. It mainly includes stationary stochastic process,

DOI: 10.54254/2755-2721/102/20241020

auto regressive process, and moving average process. In the short-term passenger flow prediction of one of the metro stations in Shen Zhen, China, the ARIMA model was also chosen.

3. Results and discussion

3.1. Preliminary work

As it's pointed in the "Method" part, the Augmented Dickey-Fuller (ADF) test was made firstly (table 1). In this process, different difference orders (from 0 to 2) are tried. Finally, a value of 2 for the difference order is confirmed to be optimal. As can be seen from Table 1, the p-value of this set of data is 0, which is less than 0.05. Therefore, this set of data is relatively stable and can be predicted using the ARIMA model. However, the ADF test alone cannot confirm the high feasibility of the ARIMA model. For that reason, another crucial analysis is done, and that is the partial (auto)correlation analysis.

Table 1. Passenger flow- Augmented Dickey-Fuller test Form

Differential order	t	p	Threshold		
			1%	5%	10%
2	-14.469	0.000	-3.435	-2.864	-2.568

It is easy to see from Figure 1 that the image of the Auto correlation Function (ACF)-Lag order is convergent, indicating that the ARIMA model can be used on this data set. Of course, ACF is not the only one that can help illustrate this point. The representation of PACF images can say the same thing.

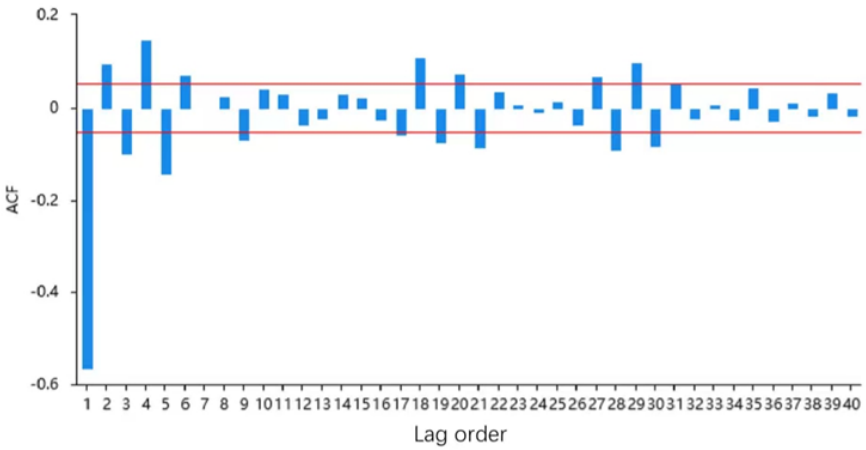

Figure 1. Auto correlation Function (ACF)-Lag order

DOI: 10.54254/2755-2721/102/20241020

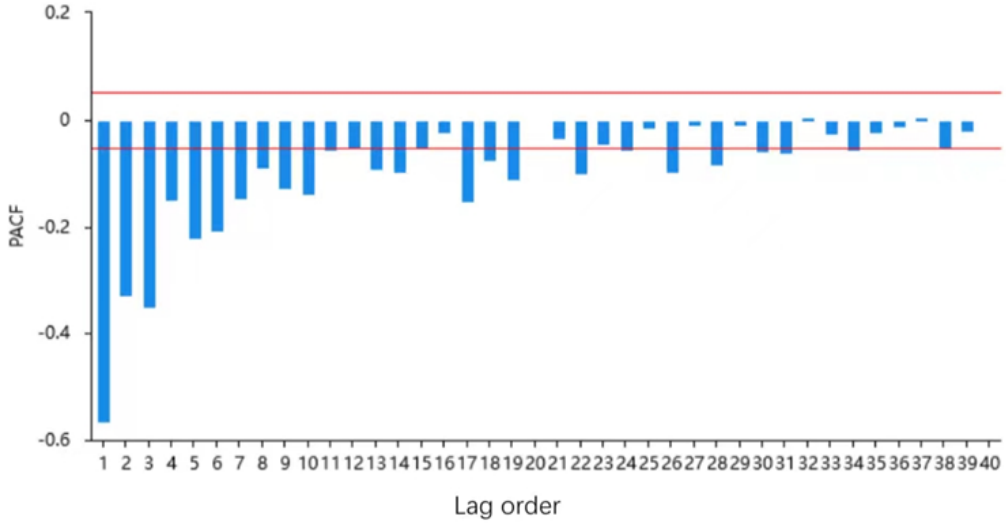

Figure 2. Partial Auto correlation Function (PACF)-Lag order

In the ACF plot (Figure 1), since there are only 40 orders of data, it is unknown whether the image will be truncated in the subsequent process. However, in PACF images (Figure 2), the truncation phenomenon at the 40th order is easily observed. With the dual support of ACF images and PACF images, the ARIMA model was finally adopted into this paper.

3.2. Model results

Then came the application of the ARIMA model. In order to simplify the application of the ARIMA model, the ARIMA model analysis in this article is obtained from the SPSSAU website. The data is fed to the SPSSAU website and analyzed and predicted using the ARIMA model. Initially, the ARIMA model (2,2,2) was adopted, i.e., the auto regressive order p, the moving average order q, and the difference order d were all 2. However, the results of this did not pass the white noise test. That is, the residuals of the model compared to the true values are not white noise. As the Figure 3, the attempt at ARIMA (2,2,2) was unsuccessful. In Figure 3, the line of predicted value deviates far from that of the true value.

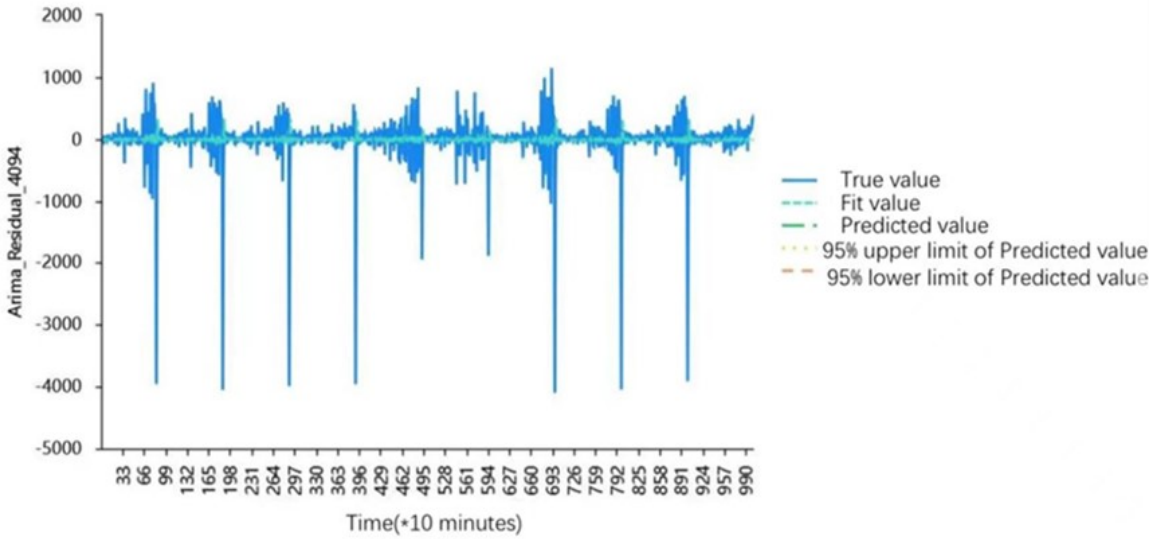

Figure 3. The result of prediction based on the model ARIMA (2,2,2)

DOI: 10.54254/2755-2721/102/20241020

It is of great importance to perform stationary detection and white noise detection on the training set before using the ARIMA algorithm [13]. The reason is that the algorithm is only suitable for time series with stable characteristics and in accordance with the white noise principle. From this point of view, white noise detection of the results is also crucial. The method is as follows: first the author performs white noise detection, and then compare the value of the white noise result parameter with the magnitude of 0.5. If the value is less than 0.5, the previous assumption is correct within the 95% confidence level interval. This allows the previous assumptions to be correct within the 95% confidence level interval. It is found that the p value of Q6 is greater than 0.1, which indicates that the null hypothesis cannot be rejected at the significance level of 0.1, and the residual of the model is white noise, and the model basically meets the requirements (Figure 4). Additionally, the mean square error (MSE) of the prediction also decreased from 165656.237 to 154638.772, which also proved the advance of the new model.

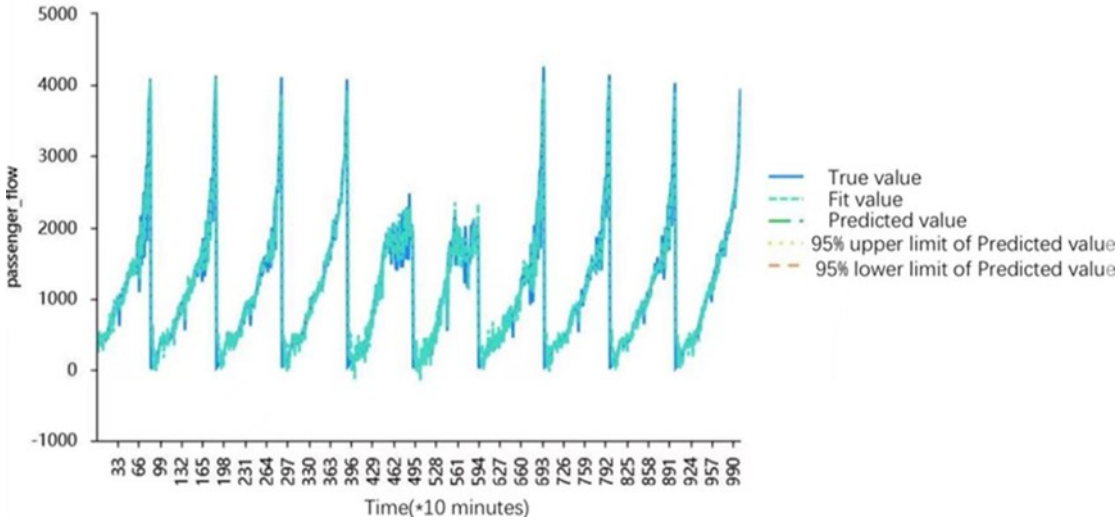

Figure 4. ARIMA Model (12,2,12) results

3.3. Discussion

In the end, the ARIMA (12, 2, 12) model is considered to be one of the more representative and reliable results displayed in this Paper. It can be seen from Figure 4 that the intersection of the predicted value line and the real value line in the prediction image using this model is more suitable in the prediction image of the ARIMA (2, 2, 2) model. And it can be clearly seen that the prediction model will be more effective in fitting for periods with less passenger flow fluctuations.

Although the model cannot accurately predict the arrival of the peak of passenger flow during the period when the passenger flow fluctuates greatly, it is not difficult to see that the model has been able to estimate the peak passenger flow more accurately by the highly coincide between the predicted value map line near the peak and the real value map line on the image enclosed with the time axis. This result is highly consistent with previous studies. In terms of realizing the function of predicting peak passenger flow, the model used in this Paper is enough to achieve similar results with other models. If you want to continue to optimize the model on this basis, you must use a larger database and more accurate data for processing. Unfortunately, due to permission restrictions, objective technical problems and cost problems, this article also has to use this data set provided by Alibaba when constructing the model. Because the data set is more detailed in time granularity than some time span monthly data sets; and compared with some data sets that are accurate to one minute, its number of samples is enough. Therefore, this appropriate data set was selected. But in any case, the results show that the ARIMA model is quite effective in short-term passenger flow prediction.

DOI: 10.54254/2755-2721/102/20241020

4. Conclusion

By observing the short-time passenger flow predicted by the ARIMA model, the obvious result is that the ARIMA model can predict the short-time passenger flow of the metro under reasonable parameter settings. In the results obtained, it was found that the predicted value fit well with the true value. This not only shows that the idea of using the ARIMA model to predict the short-time passenger flow of the metro proposed at the beginning of the paper has been realized, but also reveals the potential of the application of the ARIMA model in other aspects, such as buses, shopping malls and other scenarios. At the same time, this result also illustrates the importance of the rationality of parameter setting to the accuracy of the ARIMA model for short-term passenger flow prediction. This means that more reasonable parameter settings can often bring higher prediction accuracy. In addition, it is also very important to choose a suitable ARIMA model or independently build a new and appropriate ARIMA model when forecasting the short-term passenger flow of different scenarios in the future. It is also because of the strong adaptability of the ARIMA model that this paper believes that the ARIMA model will play a role in the field of short-term passenger flow prediction in the future, despite the emergence of models such as LSTM. The fact is that the ARIMA model has been continuously improved and even integrated with other models in recent years, so as to have a better predictive effect.

This paper adopts non-programmable data processing methods to use the ARIMA model to process short-term passenger flow data of the subway, which means that with the emergence of data processing websites such as SPSSAU, the threshold for people to process data is being lowered. This result can inspire those who have not learned programming to use similar tools to utilize big data. In this way, big data will have a wider scope of application and a deeper mass base. For those who are eager to obtain useful information from big data, such as some government civilian officials and the sales departments of some enterprises, using the ARIMA model for analysis with the help of similar websites will make their work more efficient. Finally, it must be pointed out that the method used in this paper also has its shortcomings. That is, the prediction model can only accurately predict the flow of short-term passenger flow, but it cannot accurately predict when the corresponding passenger flow will arrive. This also means that the predicted peak passenger flow often has a few minutes of deviation from the actual value at the corresponding moment. If the prediction accuracy can only reach this point, the prediction results will somewhat lose the preventive effect of short-term excessive passenger flow. However, in the future, with the increase in data collection and the improvement of the ARIMA model, the accuracy of the ARIMA model for short-term passenger flow prediction will be further improved.

Authors contribution

All the authors contributed equally and their names were listed in alphabetical order.

References

[1] China Urban Rail Transit Association 2023 Annual Statistics and Analysis Report of Urban Rail Transit. Beijing: China Urban Rail Transit Association.

[2] Yang C, Ya S, Li L J, Guo X Q and Zhang W 2024 Research on Short-term Passenger Flow Prediction of Urban Rail Transit Based on CNN-LSTM. Working paper, 11, 201-233.

[3] Xue Q, Zhang W and Ding M 2023 Passenger flow forecasting approaches for urban rail transit: a survey. International Journal of General Systems, 52(8), 919- 947.

[4] Liang S, Ma M, He S and Zhang H 2019 Short-Term Passenger Flow Prediction in Urban Public Transport: Kalman Filtering Combined K-Nearest Neighbor Approach. IEEE Access 7, 120937-120949.

[5] Liu S Y, et al. 2021 Research on Forecast of Rail Traffic Flow Based on ARIMA Model. Phys. Conf. Ser. 1792, 12065.

[6] Smith B L, Williams B M and Oswald R K 2002 Comparison of parametric and nonparametric models for traffic flow forecasting. Transp. Res. Part C Emerg. Technol, 10, 303-321.

[7] Yang J and Hou Z S 2013 A grey Markov based on large passenger flow real-time prediction model. Beijing Jiaotong Univ, 37, 119-123.

DOI: 10.54254/2755-2721/102/20241020

[8] Wang Q W, Chen Y R and Liu Y C 2021 Short-term passenger flow prediction of urban rail transit based on convolution long short-term memory neural network. Control and Decision, 36(11), 2760-2770.
[9] Zhang J, Chen F and Cui Z 2021 Deep Learning Architecture for Short-Term Passenger Flow Forecasting in Urban Rail Transit. IEEE Transactions on Intelligent Transportation Systems, 22(11), 7004-7014.
[10] Zhang T 2023 Research on short-term passenger flow prediction of urban rail transit based on TCN-LSTM combined model. Beijing: Beijing University of Chemical Technology.
[11] Li Z N 2023 Research on short-time passenger flow forecasting method for inbound (outbound) station of urban rail transit. Control and Decision, 19, 158-141
[12] Li X, et al. 2024 Research on short-term passenger flow prediction of scenic spots based on improved Transformer model. Working paper, 10, 213-224.
[13] Zhang G Y and Jin H 2023 Research on the prediction of short term passenger flow of urban rail transit based on improved ARIMA model. Control and Decision.