Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 5
AIDML 2024

Prediction of Medium-duration Subway Passenger Flow Volume based on the ARIMA Model

Mingyue Yang*

Department of Transportation, Dalian Jiaotong University, Dalian, 116028, China

* Corresponding Author Email: 100556@yzpc.edu.cn

Abstract. The subway industry has brought about a significant change in transportation by effectively reducing ground traffic congestion, resulting in an increasing demand for predicting subway passenger flow based on historical data. Researchers are constantly striving to improve the diversity and accuracy of data prediction models. This paper examines the daily passenger flow data of Nanjing Metro in 2023 and makes medium-term predictions using the ARIMA model to explore the feasibility and effectiveness of this approach. ADF, ACF, and PACF tests are conducted on the data to ensure that the parameters input into the model can optimize its accuracy. Then the ARIMA model is utilized to fit the data, resulting in a highly accurate parameter model. The results predict the passenger flow of the Nanjing subway in the next ten days, showing that the model's fitted values closely resemble the true values' distribution. By utilizing the ARIMA model, predictions for the next 10 days are made, yielding relatively accurate results. This paper demonstrates that the ARIMA model can be effectively applied to predict subway passenger flow in the medium term, thereby improving the precision of the forecasts.

Keywords: ARIMA; metro; passenger flow forecasting.

1. Introduction

Since the onset of the global industrial revolution, the influx and diversity of urban transportation options have caused a significant increase in traffic congestion. The subway, a remarkable technological feat, has been instrumental in reducing on-the-ground traffic congestion. As China's urban rail transit improves by leaps and bounds, the passenger flow in urban subways has been increasing daily. Consequently, the need for rational control of subways based on reliable data has become apparent. Predicting passenger flow is an essential aspect of studying urban rail transit networks and serves as a foundation for passenger transportation organizations. Insights gained from such investigations can inform operational management and emergency response decisions. The forecasting of passenger flow, coupled with the design and overall construction of the city, can further enhance the overall performance of rail transit. Given the current advancements in science and technology, accurately describing and forecasting passenger flow remains a highly complex and challenging task.

Various methods have been employed for passenger flow prediction, including regression analysis, cross-classification analysis, and others. For instance, Yang et al. utilized the K-nearest nonparametric regression forecasting method to construct a prediction model for inbound and outbound station volume to analyze the passenger flow prediction of Guangzhou Metro [1]. Ji integrated the growth rate method and gravity model method with the super network to forecast the passenger flow of the Wuhu light rail [2]. Meanwhile, Guang et al. applied fuzzy clustering to predict inbound and outbound traffic without relying on specific economic data [3]. Despite their usefulness, these methods have limitations, as their accuracy is influenced by subjective judgment and other factors, and their flexibility is limited. To address these concerns, Long used deep learning methods to study passenger flow prediction [4]. Additionally, Pan compared the Autoregressive Integrated Moving Average Model (ARIMA) and the Long Short Term Memory model (LSTM) and concluded that ARIMA exhibits superior prediction performance because the root-mean-square error of ARIMA is smaller than that of LSTM [5]. However, it is worth noting that most existing studies have focused

on short-term prediction (10-60 minutes) using the ARIMA model [6]. Therefore, medium-term passenger flow prediction requires further research and attention, specifically the analysis and forecasting of subway passenger flow daily (24 hours).

The ARIMA model operates on the principle of leveraging past data to anticipate future events. To achieve this, the sequence is first differentiated and transformed into the Autoregressive Moving Average Model (ARMA) model [7-9]. A trial model is then identified and diagnosed, with necessary adjustments made in a repeated cycle of identification, estimation, and diagnosis until a suitable model is established. The resulting model is a combination of an autoregressive and a moving average model, or ARIMA. Presently, ARIMA models based on Spark, ARIMA models combined with SPSS, and other methods are being employed to predict subway passenger flow [10, 11].

Therefore, this paper aims to analyze passenger flow in and out of urban rail transit stations through time series analysis based on historical data. Additionally, this study examines the accuracy and feasibility of the ARIMA model, serving as a foundational step toward enhancing subway operation modes.

2. Methods

2.1. Data Source

This paper utilizes objective passenger flow data obtained from Nanjing Metro, covering a period from January 1 to December 31, 2023, stored in CSV files. The data was collected at 24-hour intervals, revealing a daily passenger flow of approximately 3.3 million individuals utilizing the Nanjing Metro system.

2.2. Indicator Selection and Description

Table 1 presents a complete overview of the three variables used in the study, including their full names, explanations, and respective quantities. The data has been sourced from Nanjing Metro's daily passenger flow information, which is publicly available on Weibo. It's worth noting that the study's results are reliable, given the high accuracy rate of the data counts.

Table 1. Name and explanation of variables

Full Name	Explanations	Amount
Date of statistics	Inbound time	365
Passenger flow	Number of passengers entering the station	365
Data sources	@ Nanjing Metro	

2.3. Method Introduction

The ARIMA model is a statistical model for time series data analysis and prediction. This model primarily focuses on predicting the delay magnitude caused by the dependent variables, estimating the delay magnitude after accounting for model-induced uncertainty bias, and assessing the current level of modeling as the time series gradually attains stability. The ARIMA model encompasses three components, namely the autoregressive (AR), integrated (I), and moving average (MA) parts, which collectively contribute to its name.

The ARIMA (p, d, q) model consists of three parts: "p" represents the autoregressive part. This section explains the lag values of the observations used in the model. A fundamental premise of autoregressive modeling is that each observation is a linear combination of its preceding p values. The specific mathematical form is:

$$AR(p): y(t) = c + \varphi_1 * y(t-1) + \ldots + \varphi_p * y(t-p) + \varepsilon(t) \quad (1)$$

Among them, $\varphi_1 \ldots \varphi_p$ is the autoregression coefficient, c is a constant.

Where “q” stands for “moving average part”: this part describes the lag value of the error item used in the model. The moving average model establishes a relationship between the current value and the past white noise. The specific mathematical form is:

$$MA(q): y(t) = c + \theta_1 * \varepsilon(t-1) + \ldots + \theta_p * \varepsilon(t-q) + \varepsilon(t) \quad (2)$$

Among them, $\varepsilon(t)$ is the white noise error term, and c is a constant.

Where “d” is the order of the difference. The goal of difference is to transform non-stationary series into stationary series. If it is denoted as the difference operator, then here is:

$$\nabla^2 y_t = \nabla(y_t - y_{t-1}) = y_t - 2y_{t-1} + y_{t-2} \quad (3)$$

Setting up an ARIMA model involves several steps, such as determining the values of p, d, and q, estimating parameters, and checking the model's accuracy. With the help of this model, this paper can forecast and analyze future time series data.

3. Results and Discussion

3.1. Time Series Plot

Figure 1 presents a time-series chart illustrating the passenger volume of the Nanjing Metro. Based on the observations from the chart, it can be concluded that the passenger flow volume of the Nanjing Metro experiences irregular variations but exhibits periodicity overall. Therefore, an ARIMA model should be established to predict the passenger flow.

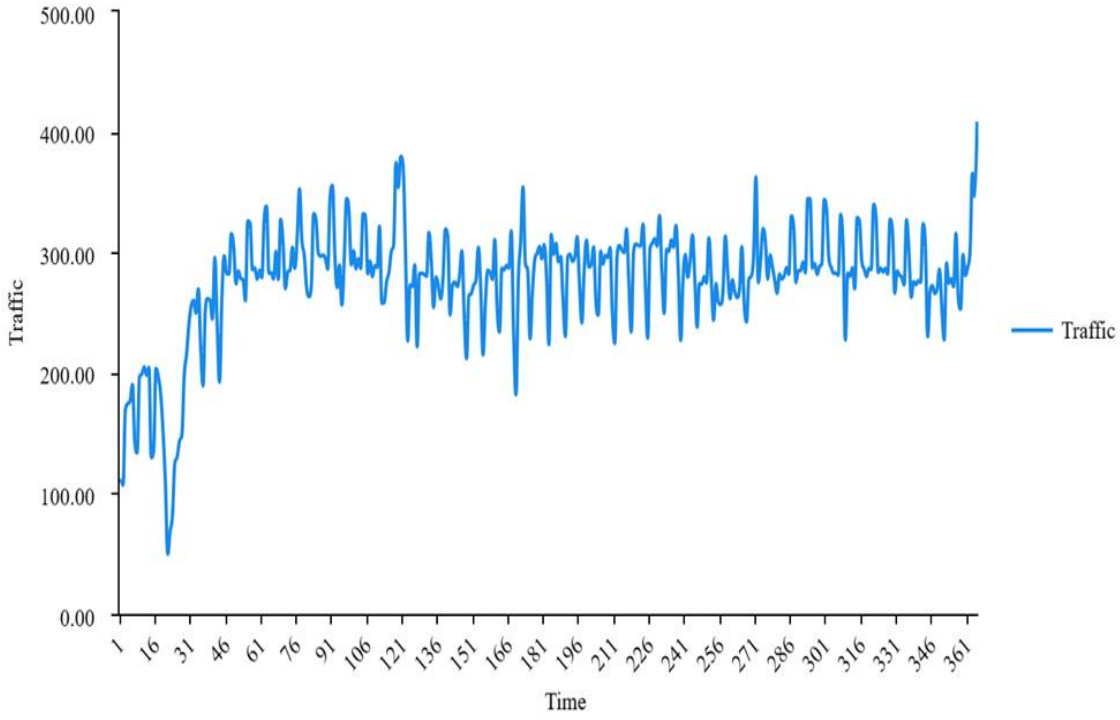

Fig. 1 Passenger flow data of Nanjing Metro in 2023

3.2. Stationarity Detection

Based on the findings outlined in Table 2, the t-statistic for the ADF test applied to the time series data is -2.925, with a corresponding p-value of 0.043. The critical values for 1%, 5%, and 10% are -3.449, -2.870, and -2.571, respectively. Given that the p-value is less than 0.05, indicating a high level of confidence (beyond 95%), it can be rightfully rejected the null hypothesis and concluded that the series is indeed stable.

Table 2. ADF test

Differencing Order	t	p	Critical Value		
			1%	5%	10%
0	-2.925	0.043	-3.449	-2.870	-2.571
1	-6.196	0	-3.449	-2.870	-2.571

ACF and PACF graphs can be used to determine the autoregressive order p and the moving average order q. As shown in Figure 2 and 3, both the ACF and PACF plots exhibit significant non-truncation, necessitating the selection of an appropriate ARIMA model order. Further analysis of Figures 3 and 4 reveals that the optimal q value is 0 and the optimal p value is 2.

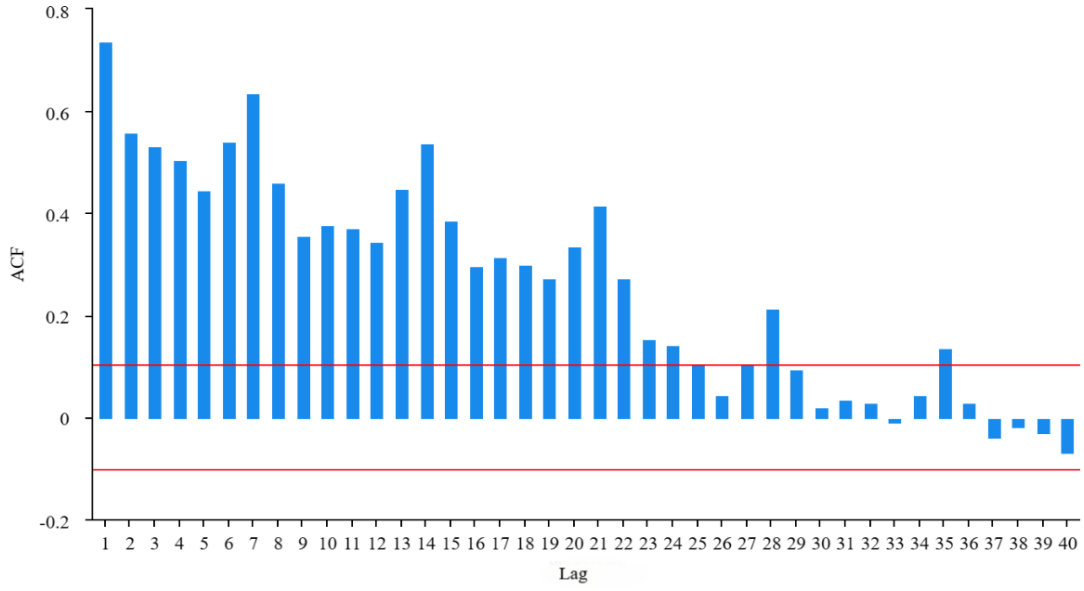

Fig. 2 ACF plot

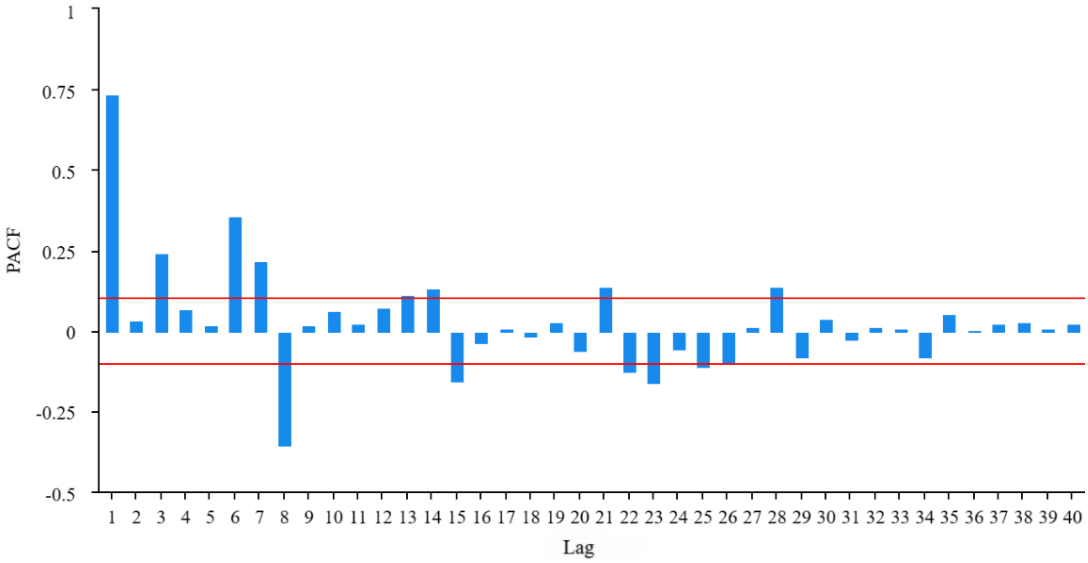

Fig. 3 PACF plot

3.3. Model Building

After constructing and comparing multiple potential candidate models, the optimal model is identified as ARIMA (0,1,2). Table 3 further compares of Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and Root Mean Square Error (RMSE) values of multiple ARIMA

models with similar parameters. The optimization process of ARIMA model construction is shown by selecting the ARIMA model with the lowest AIC, BIC, and RMSE values.

Table 3. AIC, BIC, and RMSE values

ARIMA	AIC value	BIC value	RMSE
(0,1,2)	3497.951	3513.539	29.2359
(1,1,2)	3499.471	3518.957	29.2159
(1,1,1)	3510.119	3525.708	29.7212
(0,1,1)	3551.367	3563.059	31.5306
(1,0,1)	3555.076	3570.675	31.8724

By comparison, the optimal model chosen ultimately is ARIMA (0,1,2). The parameter model is shown in Table 4, with the following model formula:

$$y(t) = 0.494 - 0.373 * \varepsilon(t-1) - 0.396 * \varepsilon(t-2) \quad (4)$$

Table 4. ARIMA (0,1,2) model parameter

Item	Sign	Coefficient	Standard Error	Z value	P value	95% CI
Constant Term	c	0.494	0.369	1.338	0.181	-0.229 ~ 1.217
MA Parameter	β1	-0.373	0.049	-7.595	0.000	-0.470 ~ -0.277
	β2	-0.396	0.049	-7.991	0.000	-0.493 ~ -0.298

The numerical model results in Figure 4 indicate that the model fits well and accurately predict the true value distribution. As a result, the model is suitable for predicting the daily passenger flow of the Nanjing Metro in the future.

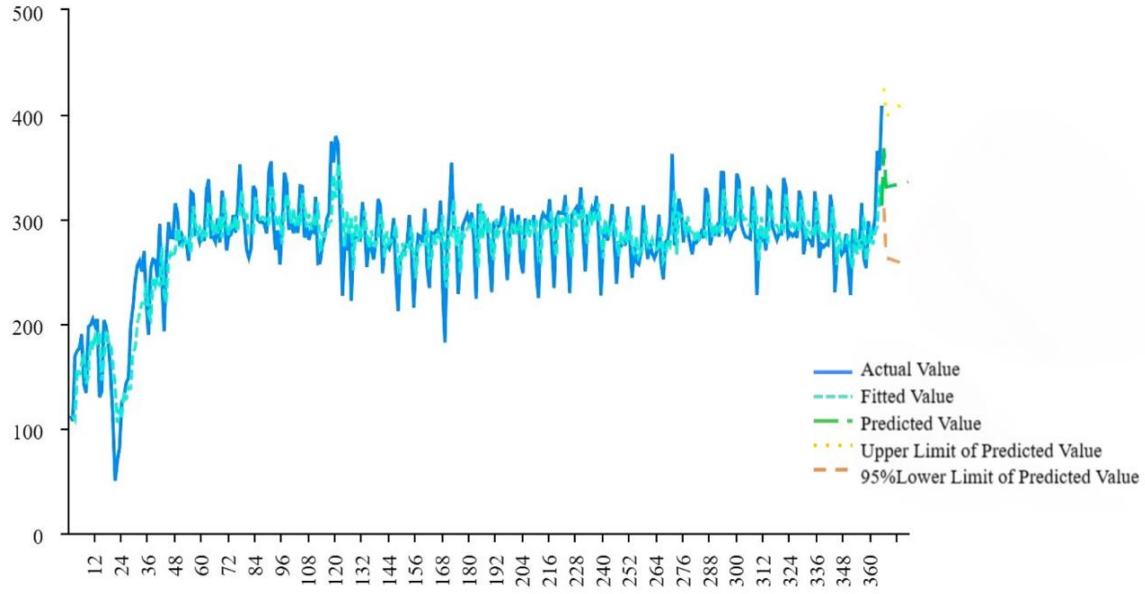

Fig. 4 Numerical model fitting and prediction

3.4. Model Prediction

The forecast for the past 10 periods is presented in Table 5. The Nanjing Metro's passenger volume is projected to decline slightly to 3.3 million over the next two days, followed by a gradual increase to 3.34 million by the tenth day. The associated metrics for this forecast are as follows: RMSE is 29.2359, Mean Absolute Error is 22.5697, Mean Square Error is 854.7357, and Mean Absolute Percentage Error is 0.0913.

Table 5. Prediction results

Prediction	Value	Prediction	Value
T=1	368.151	T=6	332.821
T=2	330.845	T=7	333.315
T=3	331.339	T=8	333.809
T=4	331.833	T=9	334.303
T=5	332.327	T=10	334.797

3.5. SARIMA Model Results

The values of PCF displayed in Figure 4 demonstrate a certain degree of periodicity, indicating that the Seasonal ARIMA (SARIMA) model could be applicable. A cycle of seven days is taken into account for the analysis. After analyzing the data, the SARIMA (0,1,1) model is utilized, and the fitting results are found to be satisfactory, as demonstrated in Figure 5. The AIC value of this model is 3357.241, the BIC value is 3376.726, and the RMSE value is 24.0972.

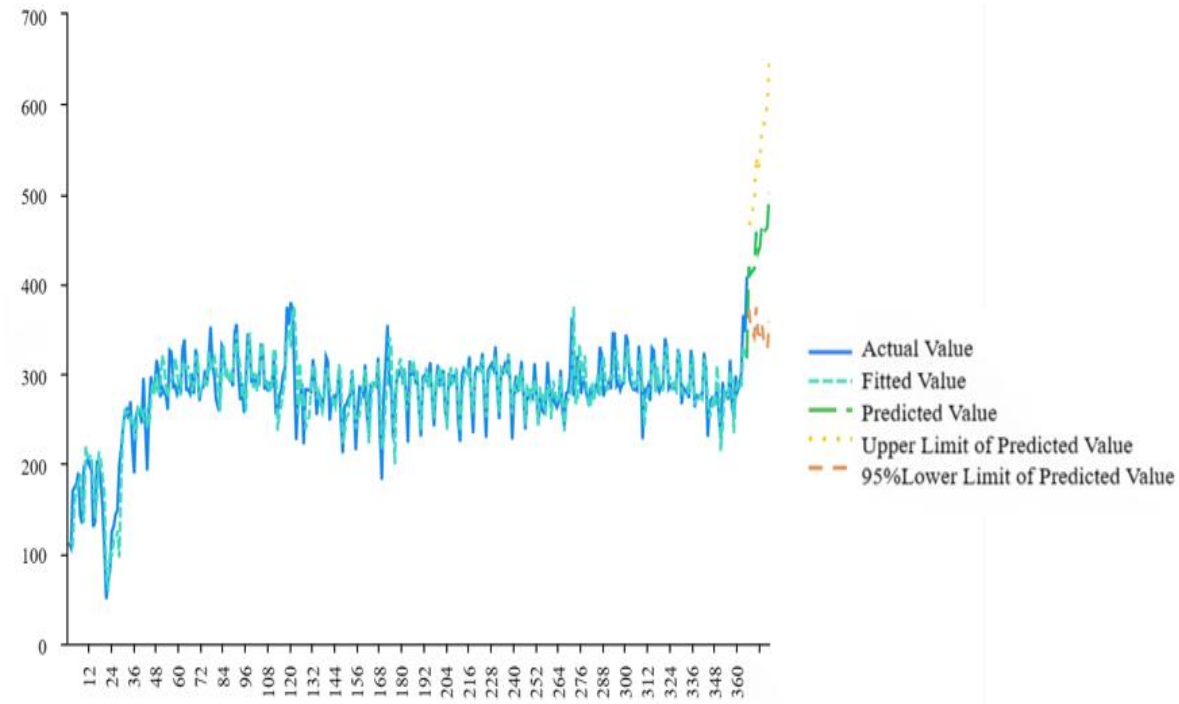

Fig. 5 SARIMA model fitting and prediction

4. Conclusion

This paper investigates the viability and efficacy of utilizing the Autoregressive Integrated Moving Average (ARIMA) model to forecast subway passenger volumes based on historical data. Through the analysis and modeling of subway passenger flow data, the passenger flow of the Nanjing subway in the next ten days is obtained. This paper discovered that the ARIMA model possesses notable potential and reliability for forecasting subway passenger flow. To a certain extent, this offers assistance to subway operators in planning future subway schedules and routes. However, it is imperative to acknowledge the limitations of the ARIMA model and explore other methods and techniques to enhance the accuracy and stability of predictions. While the ARIMA model is a widely used technique for time series analysis, it may not always provide reliable predictions due to its inherent limitations. For instance, seasonal analysis can be integrated with the daily passenger flow data of Nanjing Metro, as discussed in the paper.

Therefore, incorporating other methods and techniques, such as machine learning algorithms and data visualization tools, could prove fruitful in improving predictions. By combining different approaches,

researchers can generate more robust insights, leading to better decision-making and improved outcomes in the future.

References

[1] Yang Zhiqiang, Shi Fengshou, Huang Junda, et al. Research on passenger flow forecasting method of new urban rail station based on land use. Urban Rapid Rail Transit, 2020, 33(2): 5.

[2] Ji Xiaohui. Passenger Flow Demand Analysis Model and Simulation of Urban Rail Transit. Railway Survey and Design, 2021.

[3] Guang Zhimi, Yao Enjian, Zhang Yongsheng. Predict urban railway station entrance and exit passenger flow based on fuzzy clustering analysis. International Conference on Railway Engineering, 2012, 70-75.

[4] Long Xiaoqiang, Li Jie, Chen Yanru. Short-term ridership prediction of urban rail transit based on Deep Learning. Control and Decision, 2019, 34(8): 1589-1600.

[5] Pan Nianran. Passenger flow prediction of urban rail transit based on ARIMA and LSTM. Science and Technology Innovation, 2022, 8: 165-168.

[6] Chen Yanli, Sha Yuwu, Zhu Xiaolin, Zhang Xiaohong. Passenger flow prediction of Shanghai Metro Line 16 based on Time Series Analysis. Operations Research and Fuzzy Theory, 2016, 6(1): 15-26.

[7] Zhang Jie, Liu Xiaoming, He Yulong, Chen Yongsheng. Railway passenger flow forecast based on Time series. Statistics and Consultation, 2008, 20-21.

[8] Pei Wu, Chen Feng, Cheng Liqin. Research on ARIMA Forecasting Technology for traffic volume time series. Shanxi Science and Technology, 2009, 1: 75-79.

[9] Qi Wei, Li Ye, Wang Zuoxin. Prediction method of seasonal ARIMA model in sparse traffic flow. Highway Traffic Science and Technology, 2014, 31(4): 130-135.

[10] Chang Guozhen, Zhang Qiandeng. Passenger flow prediction of urban rail transit based on Product ARIMA model. Journal of Beijing Jiaotong University, 2014, 38(2): 135-140.

[11] Guo Jiepeng, Zhong Hongbin. Implementation of urban rail transit passenger flow Prediction based on the ARIMA Model. Science and Informatization, 2019, 6: 140-140.