Research on short-term passenger flow prediction of urban rail transit based on ARIMA algorithm

Zhihan Liao*

Tianjin University of Technology and Education, Tian Jin, China

* Corresponding Author Email: 0502220210@tute.edu.cn

Abstract. The continuous growth of urban population has led to an increasing demand for urban public transport among citizens. In response, many cities have implemented urban rail transit systems to alleviate internal traffic congestion. Accurate short-term passenger flow prediction is crucial for the efficient operation of intelligent subway systems. Therefore, it is essential to establish or select a suitable model for predicting the short-term passenger flow of subways. This study utilizes ten days of travel data from the Beijing subway between May 1 and May 10, 2019. After conducting preliminary preprocessing on the 10-day subway appearance data, an ARIMA prediction model is established for initial forecasting. The accuracy of passenger volume forecast is evaluated by analyzing the variance between predicted values and actual values at specific time intervals, with results indicating that the ARIMA model demonstrates strong predictive capability.

Keywords: ARIMA, Urban rail transit, Forecast.

1. Introduction

The development of rail transit dates back to the early 19th century, when the train was one of the earliest means of rail transit. Later, the train went into the city and became the urban rail transit. In the 1960s, the world's first subway was built in London, England, and this urban rail transit that does not occupy ground space quickly spread to all over the world. Later, urban rail transit developed from metro to light rail, tram, etc., and the subway system continued to play an important role in the field of urban rail transit. It can be seen that today, with the increasing pressure of urban traffic, urban rail transit is becoming more and more popular among urban transportation modes. Because urban rail transit is driven by electric power, green and environmentally friendly, with zero emissions, and enjoys exclusive right of way, good speed and punctuality, and is safe, comfortable and saves ground space, it has gradually become an important mode of transportation to solve urban and urban commuting, and has become a core force to solve public transportation problems. However, in order to avoid blindly starting rail transit projects, the relevant documents of the Ministry of Construction of our country stipulate that the standard of urban subway traffic is to meet the one-way peak passenger flow of more than 40,000 people. Therefore, the subway system is doomed to face the challenge of significant passenger flow immediately after its completion and operation [1]. The passenger flow of urban rail transit has certain regularity, which makes it possible to predict passenger flow. Passenger flow prediction plays an important role in the design and operation management of urban rail planning. In the planning and design of rail transit system, resources can be allocated according to the results of passenger flow prediction, such as vehicle grouping and the number of equipment inside the station [2]. The analysis of the characteristics of passenger flow rules is conducive to the overall grasp of the construction and operation of urban rail transit and the understanding of passenger travel habits, so as to guide the construction and improvement of urban rail transit. The realization of short-term passenger flow prediction can obtain the passenger flow volume in the future period, help the subway to give early warning and assist evacuation during the peak period, optimize the subway operating environment, plan subway scheduling in advance, standardize ticket pricing, and formulate reasonable volume matching, etc., so as to ensure the stable operation of the subway [3].

Research on passenger flow prediction mainly includes long-term forecast and short-term forecast. Long-term passenger flow prediction is mainly applied to the infrastructure stage. By forecasting the

long-term passenger flow of lines, subway line arrangement is overall, traffic planning is planned, and project value is maximized [4]. Short-term passenger flow forecast needs a certain stability of passenger flow. In recent years, with the rapid development of the urban rail transit industry, due to the late opening and operation of the new line, the bus supporting facilities around the station are gradually complete, and passengers have become more familiar with the new line, and the original passenger flow habits will also be changed due to the convenience provided by the new line. At this time, the fluctuation of inbound and outbound passenger flow of the station tends to be stable, so historical data can be used to predict passenger flow [5]. At present, short-term forecasting is mainly applied to the operational phase. It provides convenience for citizens by predicting passenger flow changes at various stations in the short term in the future. Due to the limited number of research reports and the high requirements of computational workload, data and details, the research of short-term passenger flow forecasting has gradually increased and become more professional. In this research field, foreign research is preferred to domestic research, so China is constantly learning from foreign advanced experience, and gradually find a model suitable for long-term subway passenger flow prediction with good effect.

Based on the data of subway passengers entering and leaving the station in urban rail transit, this paper calculates the time of individual entering and leaving the station. The passenger flow of a station in a certain period of time. Based on the previous literature methods and the actual situation, ARIMA model will be used to predict the future passenger flow based on the historical passenger flow data. Then compare with the actual data to find the advantages and disadvantages of the model. Finally, by comparing the accuracy gap of different time intervals for passenger flow prediction, the most suitable time interval for prediction is found.

2. Methods

2.1. Data Source

This paper analyzes and researches the credit card data of Beijing subway from May 1 to May 10, 2019, and the data source is real and reliable. The subway station of choice is Beijing Metro Tiantongyuan North Station, which is a residential area near Tiantongyuan North Station, less affected by holidays and weekends and other factors, so it is easy to study. Table 1 shows the daily distribution data of passenger flow of Tiantongyuan North Station in 10 days.

Table 1. Daily distribution data of subway passenger flow for ten days

Date	05.01	05.02	05.03	05.04	05.05	05.06	05.07	05.08	05.09	05.10
Passenger traffic data	6444	6714	6466	6694	6806	6331	6395	6479	6729	6464

After analyzing the daily hourly ridership data, it is evident that there is a relatively low ridership during the 4:00 a.m. and 5:00 a.m. time periods, with a distinct ID format compared to other passengers. Upon further investigation, it was discovered that the Beijing subway operates from six o'clock onwards. Therefore, the data for four and five o'clock pertains to subway workers and has been excluded from our analysis, leaving only the data after six o'clock each day.

2.2. Analytical process

With n data, a sequence of ordered n real numbers is referred to as a time series [6]. The key components of a time series include temporal elements such as year, month, and day, as well as numerical elements representing specific data corresponding to the given time. Time series models can be categorized into linear stationary and nonlinear stationary types. Nonlinear stationary series are commonly known as oscillatory series, with their oscillations being attributed to the presence of noisy data [7].

The ARIMA model prediction is an analytical method based on autocorrelation among time series data. It generalizes the temporal characteristics and structure of time series data from approximately random time series data. ARIMA is a composite model comprising autoregressive (AR) processes, moving average (MA) processes, and differential algorithms. The AR model utilizes historical data preceding each timestamp to linearly scale the node's data, resulting in an autoregressive representation within the time series dataset.

$$Y_t = \mathrm{v} + \sum_{i=1}^{p} \alpha_i Y_{t-i} + \epsilon_t \quad (1)$$

Where α_i is the autocorrelation coefficient, ϵ_t is the error between the linear combination of historical and current time data, and v is the constant term.

The MA model is mainly used to reduce the errors caused by random fluctuations in the time series data, and represents the current time **t** as the accumulation of errors generated in the autoregressive process.

$$Y_t = \kappa + \sum_{i=1}^{q} \Phi_i \epsilon_{t-i} + \epsilon_t \quad (2)$$

Where, Φ_i is the coefficient after the deviation value plus the weight, q is the order linear combination of the errors generated by the extension of the current data into the autoregressive process, and **κ** is the constant term [8].

The theoretical mathematical formula of ARIMA model is as follows.

$$\Delta^d y_t = \theta_0 + \sum_{i=1}^{q} \Phi_i \Delta^d y_{t-1} + \epsilon_t + \sum_{i=1}^{q} \theta_j \epsilon_{t-i} \quad (3)$$

It can also be seen from formula (3) that the ARIMA model is essentially a linear model, which greatly limits its ability to characterize nonlinear features of time series. The modeling of ARIMA model is divided into four processes: (1) sequence stabilization processing. If the sequence is non-stationary, it can satisfy the stationarity condition model identification by difference changes. (2) Model recognition. The order **p** and **q** of the model are determined mainly by autocorrelation coefficient and partial autocorrelation coefficient. (3) Parameter estimation and model diagnosis. Estimate the parameters of the model and test the randomness test including the significance test of the parameters and the randomness test of the residual, and then judge whether the model is feasible. (4) The model with appropriate parameters is used for prediction [9].

3. Result and Discussion

The data of Beijing Metro Tiantongyuan North Station from May 1, 2019 to May 10, 2019 was selected as the training set (all the data were after 6:00 p.m. except the data from 6:00 p.m. to 8:59 p.m. on May 10), and part of the data from mid-May 1 to May 10, 2019 was used as the test set. Figure 1 shows a line table diagram of the training set. It is not difficult to see from Figure 1 that the arrangement and distribution of passenger flow in Tiantongyuan North Subway Station has certain regularity, which can be predicted.

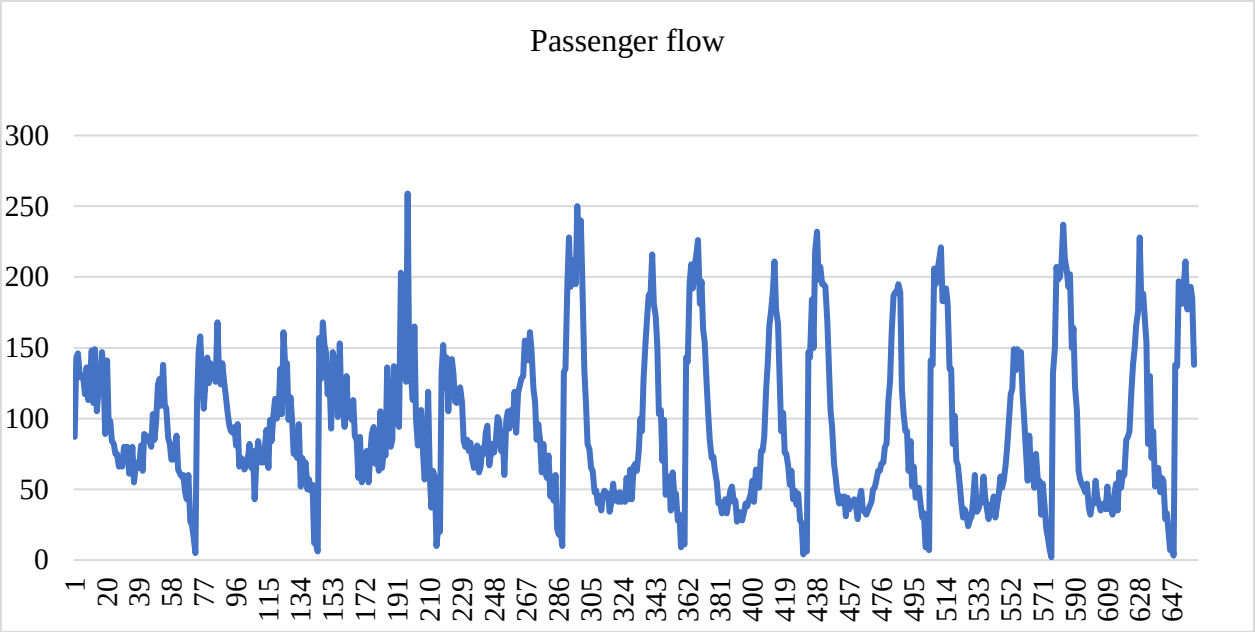

Figure 1. Statistical diagram of passenger flow data lines at different periods

ARIMA passenger flow prediction model is established based on 30 minutes time granularity data of Tiantongyuan North Station. The parameter model diagram is shown in Table 2.

Table 2. ARMA (2,2) model parameter table

term	Sign	Coefficient	Standard error	z-value	p-value	95% CI
Constant term	c	17.535	3.044	5.760	0.000	11.568 ~ 23.501
AR parameters	α_1	1.729	0.039	44.291	0.000	1.652 ~ 1.805
	α_2	-0.825	0.041	-20.109	0.000	-0.905 ~ -0.744
MA parameters	β_1	-0.690	0.058	-11.902	0.000	-0.804 ~ -0.577
	β_2	-0.261	0.058	-4.522	0.000	-0.375 ~ -0.148
AIC value: 3517.839						
BIC value: 3540.634						

Based on the passenger flow and AIC information criterion (the lower the value, the better), multiple potential alternative models were modeled and compared, and the optimal model was finally found as ARMA(2,2), whose formula was as follows:

$$y(t)=17.535+1.729*y(t-1)-0.825*y(t-2)-0.690*\varepsilon(t-1)-0.261*\varepsilon(t-2)$$

Table 3 below shows the Q statistic information of the model (specifically Ljung-Box Q test statistic), including the statistical value and data **P**.

Table 3. Statistics table for model Q

Item	Statistic	Data *p*
Q_1	0.000	0.997
Q_2	0.043	0.979
Q_3	0.587	0.899
Q_4	0.642	0.958
Q_5	1.006	0.962
Q_6	2.202	0.900
Q_7	3.191	0.867
Q_8	4.880	0.770
Q_9	10.122	0.341
Q_{10}	10.494	0.398
Q_{11}	11.237	0.424
Q_{12}	15.921	0.195
Q_{13}	17.627	0.172
Q_{14}	17.633	0.224
Q_{15}	17.736	0.277
Q_{16}	19.323	0.252
Q_{17}	21.308	0.213
Q_{18}	23.959	0.156
Q_{19}	26.702	0.112
Q_{20}	28.906	0.090
Q_{21}	28.926	0.116
Q_{22}	28.935	0.147
Q_{23}	29.212	0.173
Q_{24}	33.432	0.095
Q_{25}	33.900	0.110

The role of statistical information in model Q is as follows:

First, the arima model requires the residual of the model to be white noise, that is, the residual does not exist autocorrelation, and the white noise test can be conducted through the Q statistic test (null hypothesis: the residual is white noise);

Second, for example, Q6 is used to test whether the first 6 order autocorrelation coefficients of the residual meet white noise. Usually, if the corresponding P-value is greater than 0.1, it indicates that the white noise test is met (otherwise, it indicates that it is not white noise). In common cases, Q6 can be directly analyzed.

Third, If the white noise assumption is rejected ($p<0.05$), it means that the model is poorly fitted, and vice versa usually means that the model can be used normally.

In summary, from the results of Q statistics, if the p-value of Q6 is greater than 0.1, the null hypothesis cannot be rejected at the significance level of 0.1. The residual of the model is white noise, and the model basically meets the requirements.

Table 4 shows the predicted values and root mean square error, mean square error, mean absolute error and mean absolute percentage error for the last 10 periods.

Table 4. Predicted value (10 periods)

forecast	Backward Phase 1	Backward Phase 2	Backward Phase 3	Backward Phase 4	Backward Phase 5	Backward Phase 6	Backward Phase 7	Backward Phase 8	Backward Phase 9	Backward Phase 10
value	247.016	177.866	121.302	80.547	56.740	49.196	55.788	73.406	98.427	127.154
Root mean square error RMSE: 48.9428										
Mean square error MSE: 2395.3973										
Mean absolute error MAE: 33.0439										
Mean absolute percentage error MAPE: 0.2926										

The predicted value of the last 10 periods is compared with the real value on May 10, 2019, as shown in Figure 2.

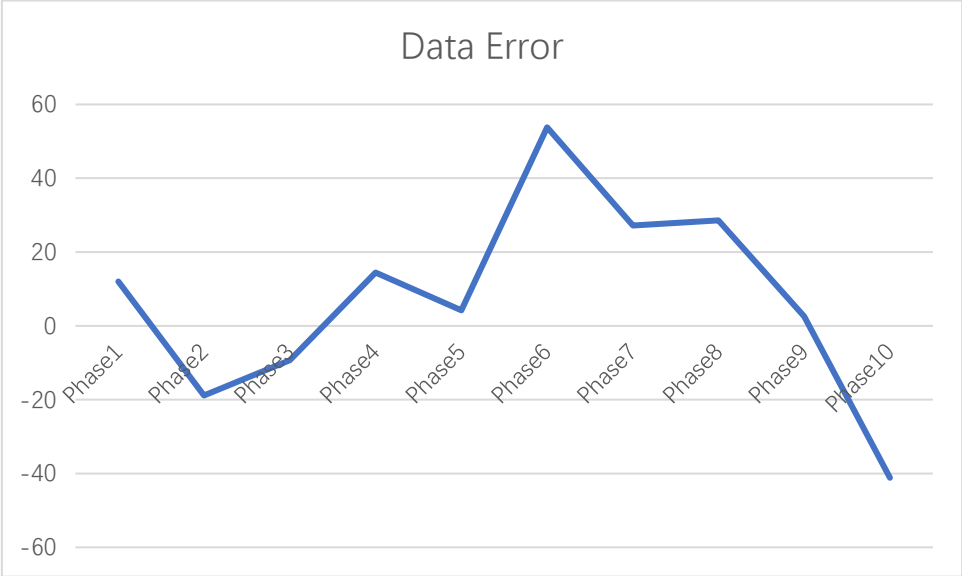

Figure 2. Error analysis

As shown in the line chart, the maximum absolute value of the error between the predicted value and the true value is 53.804, and the average absolute value of the error between the predicted value and the true value is 20.5202.

Through data analysis and statistics, it is found that the absolute error between the predicted value and the true value of the ARIMA model is small, and the mean value of the absolute error is 20.5202 and the standard deviation is 15.9387841.

Based on error data and residual analysis of model Q, it can be concluded that ARIMA model has better prediction effect and smaller error. ARIMA model can be used to predict the passenger flow of urban rail transit.

The next experiment was to investigate whether a smaller time granularity could improve prediction accuracy by reducing the time granularity to 15 minutes. Table 5 shows the predicted values, root mean square error, mean square error, mean absolute error and mean absolute percentage error for the last 10 periods.

Table 5. Predicted value (10 periods)

Forecast	Backward Phase 1	Backward Phase 2	Backward Phase 3	Backward Phase 4	Backward Phase 5	Backward Phase 6	Backward Phase 7	Backward Phase 8	Backward Phase 9	Backward Phase10
Value	132.552	108.065	97.676	76.498	68.360	52.155	47.818	37.379	37.459	32.721
Root mean square error RMSE: 24.0539										
Mean square error MSE: 578.5907										
Mean absolute error MAE: 16.6803										
Mean absolute percentage error MAPE: 0.3046										

The projected values for the next 10 periods are compared with the actual values as of May 10, 2019, as shown in Figure 3. As shown in the discounted statistical chart, the maximum absolute error between the predicted value and the true value is 27.676, and the average absolute error between the predicted value and the true value is 2.0763.

Through data analysis and statistics, it is found that the absolute error between the predicted value and the true value of ARIMA model is relatively small, the average absolute error value is 2.0763, and the standard deviation is 11.68162681.

The average value of the absolute error of the model with a time granularity of 30 minutes is 18.44 higher than that with a time granularity of 15 minutes, which reduces the error by about 900%.

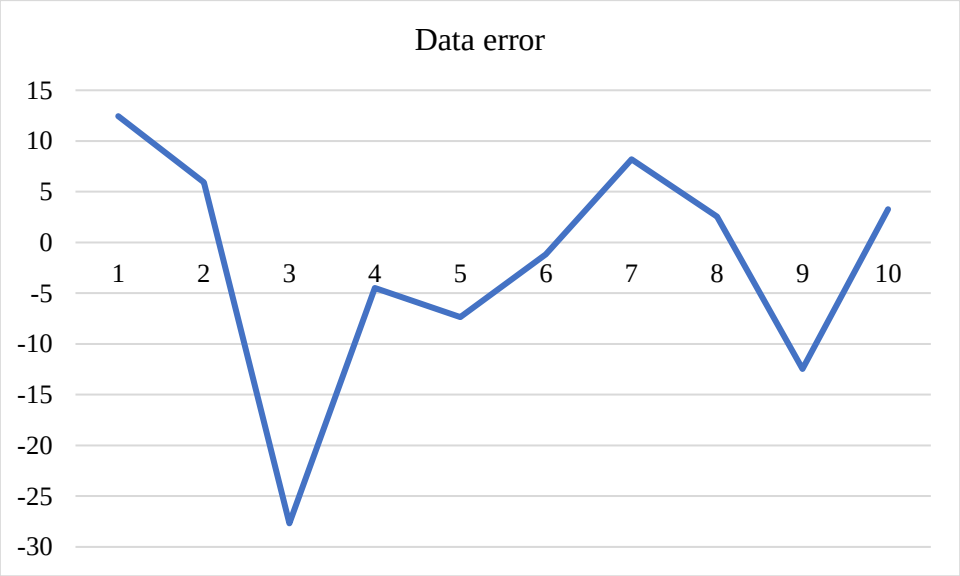

Figure 3. Error analysis

4. Conclusion

From the error data and residual analysis of model Q, the following conclusions can be drawn: ARIMA model has good prediction performance and small error, and can be used to predict the passenger flow of urban rail transit. In addition, when the time granularity is reduced, the accuracy is significantly improved. At the same time, however, more accurate model results cannot be obtained due to the limited number of days in the sample set, which makes the long-term forecast results inconsistent with the actual situation in the future.

At the same time, unexpected factors such as holidays, fares and other modes of transportation should be considered in the future model prediction, which will make the prediction result more accurate and have certain practical significance for urban rail transit operators.

References

[1] Hou Xiufang, Mei Jianping, Zuo Chao et al. Statistical Analysis of urban rail transit lines in 2020 [J]. Urban Rapid Transit, 21,34(03):1-9+64.

[2] Lu Zhongshi. Intersectional operation optimization of urban rail trains based on short-time passenger flow prediction [D]. Hefei university of technology, 2022. DOI: 10.27101 /, dc nki. Ghfgu. 2022.001454 pp1-2.

[3] Zhang Liwen. Research on short-term passenger flow prediction of rail transit based on big Data [D]. Beijing university of chemical industry, 2023. DOI: 10.26939 /, dc nki. Gbhgu. 2022.000969. P3.

[4] Liu Jingya. Analysis of passenger flow characteristics and short-term prediction of urban rail transit based on AFC data [D]. Beijing jiaotong university, 2023. DOI: 10.26944 /, dc nki. Gbfju. 2022.002194.

[5] Chen Xiaojian, Tang Qiusheng. Research on passenger flow prediction of Metro network based on multi-mode grey model [J]. Transportation Science and Economics,2019,21(04):16-20.

[6] Liang Ke. Analysis and prediction of passenger flow characteristics of urban rail transit based on ARMA model [D]. Shanghai university of engineering science, 2021. DOI: 10.27715 /, dc nki. GSHGJ. 2020.000394.

[7] He Xiaoxu. Research on some Key Problems in Time series data Mining [D]. University of Science and Technology of China,2014.

[8] Zhao Xiaoli. Application of ARIMA Model in short-term passenger flow prediction of urban rail transit [J]. Modern Urban rail Transit,2023(08):77-82.

[9] Xiong Zhibin. Research on GDP time series forecasting based on ARIMA and neural network integration [J]. Mathematical statistics and management, 2011, 30 (02) : 306-314. The DOI: 10.13860 / j.carol carroll nki SLTJ. 2011.02.012.