Proceedings of the 2nd International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/101/20241038

Enhancing periodic mobility planning: ARIMA-driven prognostications for subway passenger volume dynamics

Minxuan Zhang

School of Traffic and Transportation, Lanzhou Jiaotong university, Gansu, 730070, China

20220101548@stu.lzjtu.edu.cn

Abstract. This study aims to accurately predict urban subway passenger flow using the Auto-Regressive Integrated Moving Average (ARIMA) and Seasonal Auto Regressive Integrated Moving Average (SARIMA) model, with the intention of providing a scientific basis for subway operation management. By meticulously sorting and analyzing the existing traffic data of a large city's subway system, this research constructs a model that comprehensively considering the seasonal, trend, and stochastic characteristics of time series data. During the model construction process, data preprocessing is carried out, including stationarity tests, differencing, model order determination, and parameter estimation, ultimately identifying an ARIMA and SARIMA model with good fitting results, effectively identifying and forecasting characteristics such as numerical fluctuations. Additionally, this study conducts a special analysis of the prediction effects at diverse time scales, verifying the applicability and superiority of the ARIMA and SARIMA model in subway passenger flow prediction. This research not only provides a powerful tool for passenger flow prediction for subway operation management departments, aiding in the reasonable allocation of capacity and optimization of operation strategies, but also holds significant reference and application value for the passenger flow prediction and management of other urban public transportation systems.

Keywords: ARIMA model, SARIMA model, subway passenger flow, accuracy assessment, rail transit.

1. Introduction

By 2023, the total rail length of operation in Chinese mainland is 11 224. 54 km [1]. In the grand chapter of Chinese development of the productive force, the subway traffic system has gradually become the backbone of the transportation network with its unique charm and important role. Subway not only effectively alleviates the traffic pressure of the city and improves the travel experience of people, but also has a profound impact on the development of productivity and optimizes the urban layout. Standing at a new historical node, it indicates that the future transportation will be energy efficient intelligent and efficient. But the accuracy of passenger flow volume prediction is directly related to the productivity of the line, and inaccurate prediction may lead to waste of resources or insufficient services. Congestion in rail transit is a common problem during peak hours, affecting the passenger experience. The discussion on how to use known data to effectively forecast ridership of rail traffic transportation has opened a dynamic prelude, and the rail transit operation department can formulate transportation scheduling plans based on the short-term passenger flow prediction results. Then passengers can plan the optimal travel plan based on the forecast ridership results to achieve efficient, green and smart travel.

Liu used the ensemble algorithm Ada boost to integrate multiple sub-models into a prediction model, which proved that the ensemble model can effectively improve the prediction accuracy and reduce the risk of overfitting, and enhance the stability of the model compared to the single model in prediction accuracy and generalization ability [2, 3]. Wang proposed a forecasting model based on Convolutional Long Short-Term Memory (Conv LSTM), which fully considered the spatiotemporal characteristics of variables and made predictions based on spatiotemporal network images to ensure the full utilization and generalization ability of ridership characteristics. Meanwhile, the advantages of multiple models are used to improve the overall prediction performance and fault tolerance [4].

Combined with the previous research results on short-term traffic system state estimation, when the time granularity of the predicted traffic flow system is less than 5 minutes, the unpredictable factors lead to the enhancement of the time variability. Nonlinearity, and unreliability of the traffic flow, and the nonlinear model has strong applicability compared with the linear mathematical statistical model with poor prediction accuracy of small granularity [5]. Qi et al, proposed an improved Long Short-Term Memory (LSTM) algorithm that combines ensemble empirical pattern decomposition algorithm and Bayesian optimization algorithm. The error of forecasting results improved significantly [6, 7]. In particular, when special factors such as holidays are considered, Bai proposed a combined model of time series and regression analysis, and further improved it by introducing dummy variables and combining similar daily sample data to achieve high-precision solution of abnormal prediction problems [8]. Chen proposed an improved Kalman filter model based on back propagation neural network correction, which effectively solved the prediction error and divergence problems of traditional algorithms and improved the accuracy of bus short-term passenger flow prediction [9]. Cao et al. proposed that the prediction results of the short-term inbound passenger flow prediction model based on the Convolutional Neural Network Long Short-Term Memory (CNN-LSTM) combination model have advantages in various indicators in different scenarios [10]. Doe proposed that LSTM requires a great deal of computational data to train and is easy to overfitting. In comparison, ARIMA is a traditional statistical method. ARIMA models typically require less data and computational resources and can handle linear and seasonal patterns well [11].

However, these models are not universal enough to the stochastic passenger flow prediction, and may often fail to perform well on the generalization problem. Next, this paper will expand and analyze more data of the same type to refine the construction of the model. Further research should be based on the classification of various influencing elements, so that the established model and parameter settings have a better combination, and thus have more powerful universality and accuracy.

2. Methodology

2.1. Data source

The passenger flow volume is affected by a variety of reasons such as time period like weather, holidays, economic activities, transportation policies, service quality and residents' travel habits. Among which the immediate impact of morning and evening peak hours, bad weather and special events on passenger flow is particularly significant. As shown in Table 1, the following sections are shown due to the amount of data (Table 1).

To ensure the accuracy and completeness of the data, the data is first cleaned and preprocessed. A large amount of redundant data will increase memory consumption. increase costs, and reduce model quality. Incorrect data may lead to inaccurate final results, if not handled properly. For an algorithm to be effective, it must be provided with clean, accurate, and concise data.

Table 1. Traffic counting and time table

Time	Passenger flow rate (ten thousand)	Time	Passenger flow rate (ten thousand)
5:30	0	13:30	178
5:45	0	⋮	⋮
⋮	⋮	23:15	163
13:15	191	23:30	0

2.2. ARIMA and SARIMA model

Auto Regressive Integrated Moving Average (ARIMA) is a statistical model used for analyzing and forecasting time series data. It combines three models: Auto Regressive (AR), Integrated (I), and Moving Average (MA), adjusting parameters to fit different time series characteristics such as trends, seasonality, and noise, thus achieving data stationarity and predicting future values.

Seasonal Auto Regressive Integrated Moving Average (SARIMA) is an extension of ARIMA, tailored for time series data with seasonal fluctuations. It adds seasonal components to ARIMA, including Seasonal Auto Regressive (SAR), Seasonal Difference (S.I.), and Seasonal Moving Average (SMA), which can better capture and predict seasonal patterns in the data, making it suitable for time series forecasting tasks with pronounced periodic features.

The difference between ARIMA and SARIMA lies in their handling of seasonal data. ARIMA is suitable for time series data without obvious seasonal or cyclical patterns, capturing linear relationships and short-term dynamics through AR, I, and MA. In contrast, SARIMA is designed for data with seasonal components, adding seasonal parameters to model and predict seasonal fluctuations. The connection between the two is that SARIMA extends ARIMA by incorporating seasonal factors, enabling the model to handle both non-seasonal and seasonal characteristics simultaneously. Thus, providing more accurate and effective forecasting for time series with seasonal patterns.

Due to the regular increase in passenger traffic on weekends. Therefore, this paper will use two ways said above to forecast the passenger flow volume.

3. Results and discussion

3.1. Preliminary work

At first, this paper draws a time sequence with the data (Figure 1).

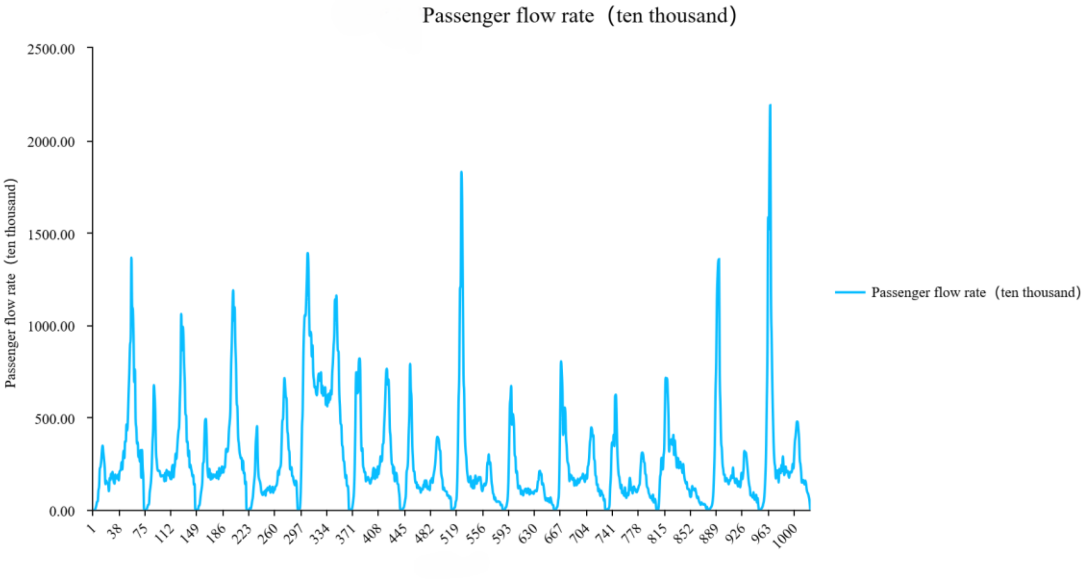

Figure 1. Time series plot

From the figure 1 and figure 2, it can be seen ridership shows a regular variation in the morning and evening peaks. On weekends, the peak will become higher. This may be because of Beijing's role as the central city of the metropolitan area. On weekends, people gather here. As a result, the number of visitors will be increased.

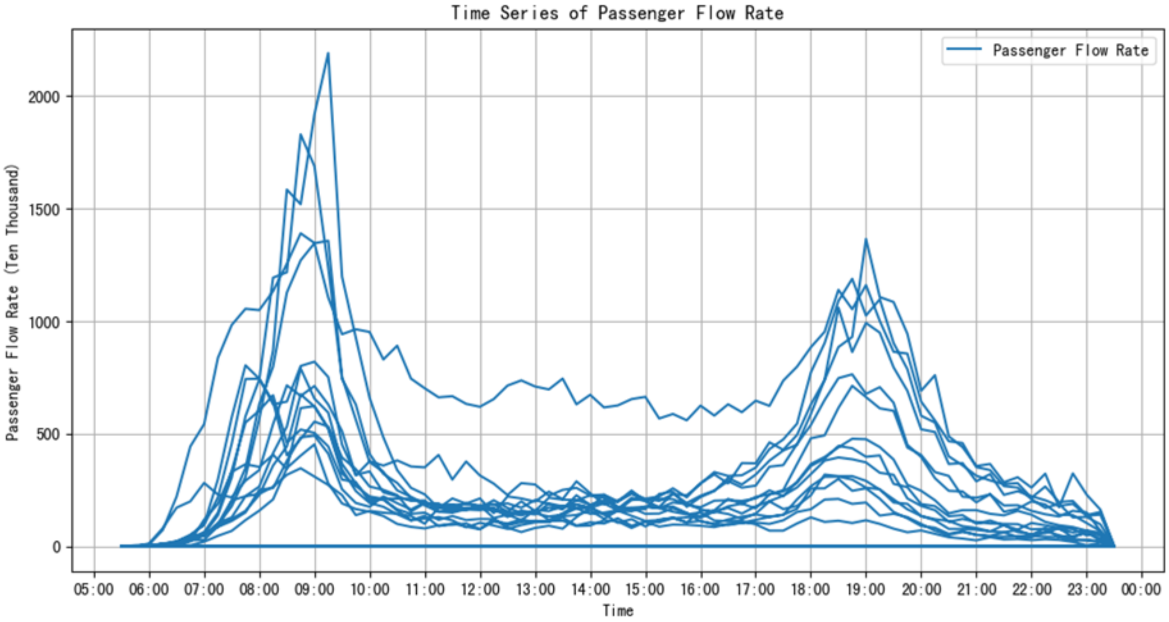

Figure 2. Time series plot of the same moment

The t test value -8. 815 is greater than all the critical values at the 1%, 5% and 10% levels, and the time series of passenger flow rate (in 10. 000 person-times) is considered to be stationary. The p-value (0. 000) is very small, which further supports the conclusion of the stationarity of the sequence (Table 2).

Table 2. ADF test.

Differential order	t	p	Threshold		
			1%	5%	10%
0	-8. 815	0. 0	-3. 437	-2. 864	-2. 568

3.2. ACF and PACF test

From the left figure in figure 3, the autocorrelation function (ACF) graph illustrates how the autocorrelation coefficient of a time series dataset varies with different lags. The coefficient is approximately 1 when the lag is 0, indicating a strong positive correlation. As the lag increases, the coefficient gradually decreases, approaching zero, which suggests that the correlation weakens over time.

On the other hand, from the right figure in figure 3, the partial autocorrelation function (PACF) graph depicts the partial autocorrelation coefficient of the same dataset at various lags. The coefficient is also around 1 at lag 0, showing a strong positive correlation. However, at lag 1, there is a notable negative value, indicating a significant inverse correlation. Beyond this point, the coefficient tends toward zero, implying that the correlation stabilizes.

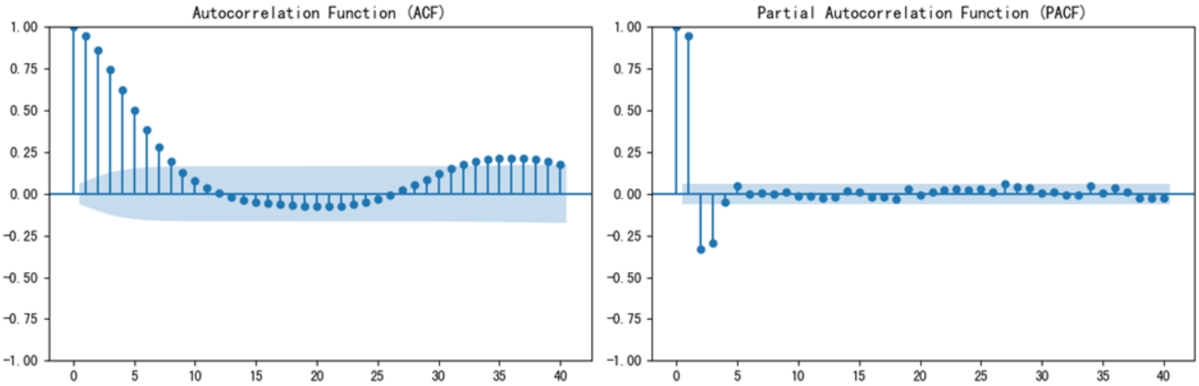

Figure 3. ACF and PACF plot.

In essence, the two graphs offer insights into the correlation structure of the database. The ACF graph reveals how the overall correlation changes with time, while the PACF graph offers a more nuanced view by isolating the correlation at each individual lag. This information can be particularly useful for identifying patterns and dependencies within the dataset, which can inform further analysis and modeling.

3.3. Model results

3.3.1. ARIMA model results

Then, ARIMA forecasts are applied (Figure 4):

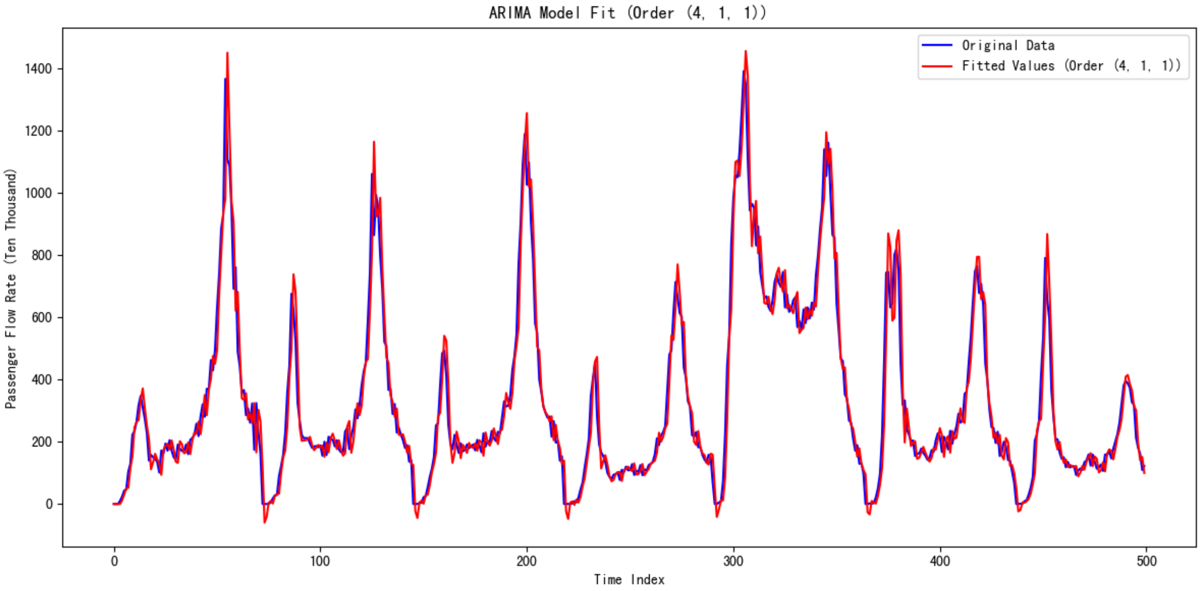

Figure 4. Forecast chart

This figure 4 exhibits the fitting effect of the ARIMA model. In figure 4, the blue line is original data and the red line is the fitted values in the ARIMA model (4,1,1). Overall, the fitted curve is very close to the original data curve, especially in most areas, the two curves almost the same trend. This indicates that the ARIMA model well captures the fluctuation pattern of the original data and fits it accurately.

In some places, such as around 200 and 400, there are differences between the fitted curve and the original data, but these differences are not large, and in the overall trend to maintain consistency. This may be because the ARIMA model is more accurate in handling short-term fluctuations, and there may be some deviation in handling long-term trends or extreme values.

Overall, this graph shows the effectiveness of the model in fitting passenger flow volume. However, different parameter combinations (p, d, q) have a great impact on the anticipation effect, and the selectivity of the parameters is obviously subjective. Therefore, this article will list six different (p, d, q) combinations to see which group works best for predictions (Table 3).

Table 3. Characteristics of different (p, d, q)

(p, d, q)	RMSE	MSE	AIC	BIC	R^2
(5, 1, 0)	83.29	6938.00	11942.29	11976.79	0.9118
(5, 1, 1)	80.57	6492.14	11877.44	11916.87	0.9173
(4, 1, 0)	83.50	6972.29	11945.29	11974.86	0.9114
(4, 1, 1)	80.66	6506.50	11877.73	11912.23	0.9175
(4, 1, 2)	80.66	6506.36	11879.70	11919.13	0.9173
(3, 1, 1)	80.78	6526.18	1187860	11908.17	0.9169

From the table 3, it can be seen (4, 1, 1) parameter combination has the lowest RMSE and MSE, and the AIC and BIC are also relatively low, so it can be considered to be the optimal parameter combination.

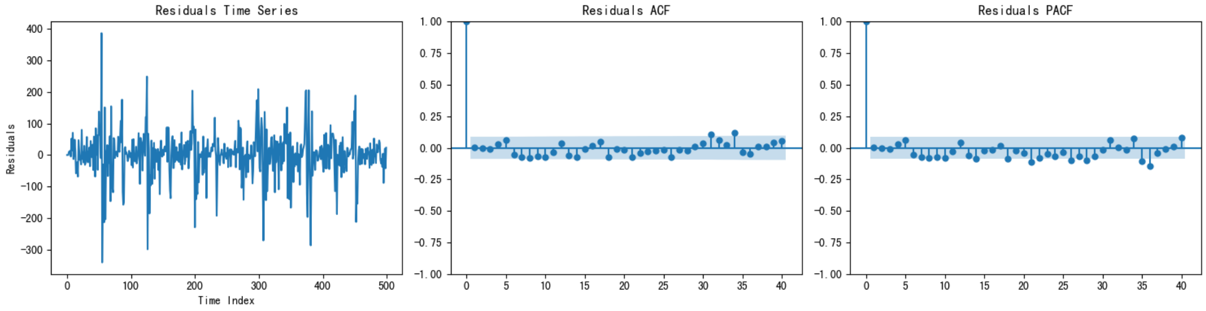

Figure 5. ARIMA (4, 1, 1) characteristics

From the figure 5, it can be seen the residual time series plot reveals a uniform distribution of residuals across the time series, showing no evident trends or periodic patterns. In the ACF plot of residuals, all autocorrelation coefficients for lags beyond the first are confined within the significance level, suggesting that there is no notable autocorrelation present among the residuals. Similarly, the residual PACF plot indicates that partial autocorrelation coefficients for lags beyond the second are also within the significance threshold, implying the absence of significant partial autocorrelation between the residuals.

These results show that the ARIMA model using the combination of (4, 1, 1) parameters fits the data well, and the residuals are distributed in white noise, indicating that the model has got most of the characteristics in the database. Therefore, this paper can consider (4, 1, 1) to be the best combination of parameters for this set of data.

On weekends, the number of visitors is regularly higher than on weekdays. Therefore, this paper will also try to drive the SARIMA model to predict.

3.3.2. SARIMA model results

Seasonal Auto Regressive Integrated Moving Average (SARIMA) is a statistical model used for analyzing and forecasting seasonal time series data. It is an extension of the ARIMA model, incorporating seasonal components to handle the periodic fluctuations present in time series data.

Among the parameters that this model needs to determine are (p, d, q), s, (P, D, Q). Where (4,1,1) is the best result compared above. Period s is one week of the data, so take 7. This paper will compare different parameter combinations to determine the best fit prediction results (Table 4).

Table 4. AIC to different (P, D, Q)

(P, D, Q)	AIC	BIC
(1,1,0)	12329	12364
(4,1,0)	12086	12136
(6,1,0)	12041	12101
(4,1,2)	11909	11968
(4,1,4)	11896	11965

This paper chose the smallest group in AIC and BIC. That is (4,1,4). The fitting prediction figure 6 is shown below.

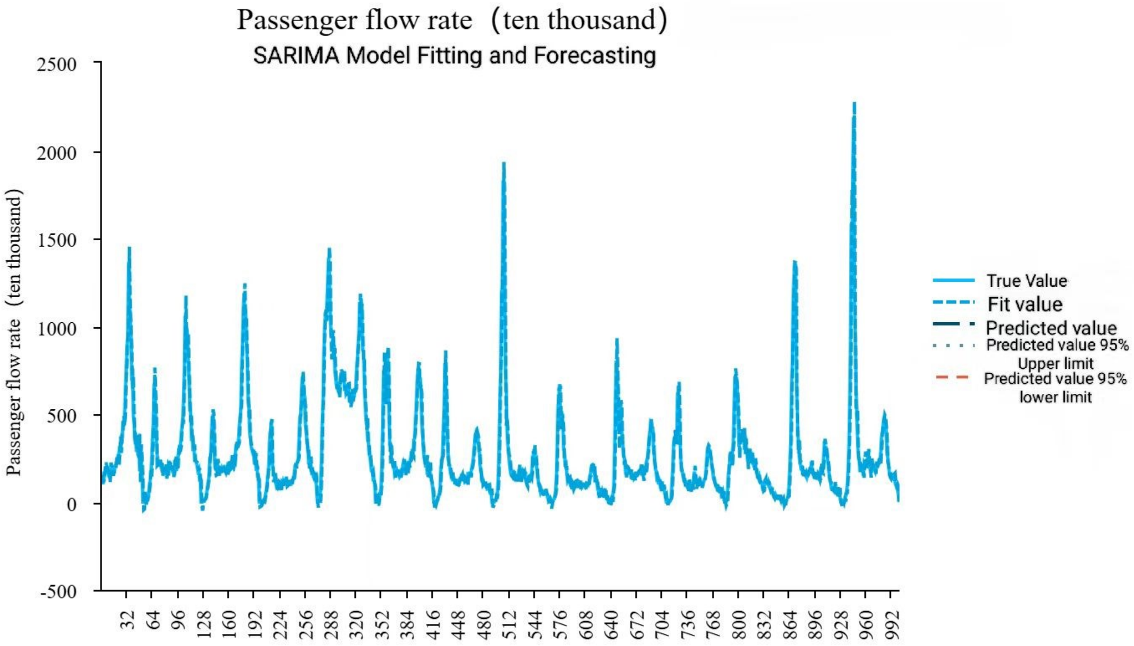

Figure 6. SARIMA (4, 1, 1),7, (4, 1, 4) Forecast chart

The residual test results of the model are shown in the table 5 below. These results show that the SARIMA model using the combination of (4, 1, 4) parameters fits the data well, and the residuals are distributed in white noise, indicating that the model has acquired most of the characteristics in the database. Therefore, this paper can consider (4, 1, 4) to be the best combination of parameters for this set of data.

At the same time, the characteristic of SARIMA model combination compared with ARIMA model combination is that it has better dynamic database capture effect during holidays. This means that the prediction is better during the holidays.

Table 5. Model residual test results

LM sequence check	number	Ljung-Box	number
F	17.697	Ljung-Box	0.038
p	0.060	p	0.846
T * R	1.782	p	0.000

Table 5. (continued)

p	0.060	JB	57420.617
H	1.553	Skew	-2.124
p	0.000	Kurtosis	39.620

4. Conclusion

This study has achieved significant research results by applying the ARIMA model to forecast subway passenger flow volume in a certain city. The ARIMA model has demonstrated good applicability and accuracy in handling time series data, effectively capturing the seasonal, trend, and random characteristics of subway passenger flow. Through the optimization of model parameters, the prediction results can accurately show the passenger flow in the future period of time, verifying the practical value of the ARIMA model in the field of subway passenger flow volume prediction. Additionally, this study provides strong decision-making support for subway operation management departments, contributing to the improvement of subway operation efficiency and passenger satisfaction.

The parameters combination of the SARIMA model has significant and far-reaching consequences on forecasting results. This study determined the parameters through ACF and PACF plots and model diagnostic indicators, but this method may involve some subjectivity. Future research can attempt to use more advanced algorithms to automatically optimize model parameters.

Consideration of external factors: Subway passenger flow is manipulated by various external factors like weather, holidays, and large-scale events. This study did not incorporate these factors into the model, which may lead to some bias in the prediction results. Future research can attempt to construct a multivariate time series prediction model that includes these external factors.

Although the ARIMA model and SARIMA model achieved good prediction results in this study, its applicability may be limited by data characteristics and the span of the prediction period. The performance of these model may vary for subway passenger flow prediction in other cities or different time periods.

In summary, this study provides an effective modeling method for subway passenger flow prediction, and it has strong applicability. However, there are still many areas worthy of further exploration and improvement. It is hoped that through subsequent research, the prediction model can be continuously refined to provide more precise decision-making support for subway operation management.

References

[1] China Urban Rail Transit Association 2024 Annual Statistics and Analysis Report on Urban Rail Transit for 2023. Beijing: China Urban Rail Transit Association.

[2] Liu J 2022 Research on Short-Term Railway Passenger Flow Forecasting Model Based on Integrated Algorithms. Journal of Chongqing Jiaotong University, 5.

[3] Zhang M J, et al. 2024 Research on Short-Term Passenger Flow Forecasting in Urban Rail Transit. Transportation Technology and Management, 5(13), 18-21.

[4] Wang Q W, Chen Y R and Liu Y C 2021 Short-Term Passenger Flow Forecasting in Urban Rail Transit Based on Convolutional Long Short-Term Memory Neural Network. Control and Decision, 11, 2760-2770.

[5] Wang H F, Teng J, Ye L, et al. 2024 Short-Term Forecasting Method of Passenger Flow Density in Urban Rail Transit Based on Nonlinear Kalman Filter. Urban Rail Transit Research, 27(6), 33-38+43.

[6] Qi X Y and Fu C H 2024 Improved LSTM Method for Short-Term Passenger Flow Forecasting in Subway. Transportation Technology and Economy, 26(02), 58-64.

[7] Zhao Q, Feng X, Zhang L, et al. Research on short-term passenger flow prediction of LSTM rail transit based on wavelet denoising[J]. Mathematics, 2023, 11(19): 4204

[8] Bai L 2017 Research on Short-Term Passenger Flow Forecasting Methods in Urban Rail Transit Under Normal and Abnormal Conditions. Journal of Transportation Systems Engineering and Information Technology, 17(1), 127-135.

[9] Chen J C 2021 Research on Short-Term Public Transportation Passenger Flow Forecasting Model Based on Improved Kalman Filter Algorithm. Dalian: Dalian Maritime University.

[10] Cao Y, Sun Y, Lin L, et al. 2024 Research on Short-Term Station Entry Passenger Flow Forecasting in Urban Rail Transit Based on CNN-LSTM. Transportation and Transportation, 40(2), 94-99.

[11] Doe J, Smith A 2020 A Comparison between ARIMA, LSTM, and GRU for Time Series Forecasting. Journal of Time Series Analysis, 45(3), 123-145.