Forecasting Passenger Flow in Chengdu Metro Based on the ARIMA Model

Yuanjin Zhu*

College of Mathematics, Sichuan University, Chengdu, China

Abstract. With the increase in population, short-term forecasting of subway traffic is becoming increasingly important. Based on data from the first half of 2025, Chengdu metro ridership, the system is constructed by using the autoregressive integrated moving average (ARIMA) model. The study first confirms the statistical properties of the passenger flow series through the Augmented Dickey-Fuller (ADF) smoothness test, and then uses autocorrelation and partial autocorrelation analyses to determine the model order, establishes the ARIMA (1,1,1)(1,0,1) seasonal prediction model, and the prediction model has a good mean square error (MSE) value of 1910.41. The prediction shows that the passenger flow will maintain the growth trend and keep the rules fluctuating. The results provide a basis for the operations department to rationalize vehicle schedules.

1 Introduction

With the acceleration of urbanization in Chengdu, the subway has become the main way for citizens to travel. Cheng pointed out that the acceleration of rail transit into a network is an important milestone in the history of the development of a modern comprehensive transportation system [1]. In the first half of 2025, Chengdu Metro has become a large-scale operation system, with the average daily passenger flow reaching the level of one million people.

However, the large passenger flow in subway stations also brings a lot of harm. Xi pointed out that the backlog of passenger flow leads to the prolongation of the train's stay in the station, resulting in extensive paralysis of the road network, and is likely to lead to overloading of some electromechanical equipment and injuries to special groups [2]. Therefore, accurate passenger flow prediction is of great significance.

In this context, the autoregressive integrated moving average (ARIMA) model has unique value in short-term forecasting scenarios. Based on it, Zhu established a pepper price forecasting system and successfully forecasted the price of small agricultural products [3]. Based on the ARIMA model, Liao predicted how the establishment of the Chengdu-Chongqing Twin Cities Economic Circle would impact Chongqing residents' consumption and the time for residents' consumption to return to normal after five years [4]. Wang made a short-term prediction of the Wuhan-Hamming Intercity Railway and obtained an average error of 7% in the model prediction, which demonstrated the high reliability of the ARIMA model in short-term passenger flow prediction [5].

Therefore, in this paper, the ARIMA model is used to predict the passenger flow of Chengdu Metro; the hope is to use historical data to get an accurate prediction for the short term.

2 Methods

2.1 Data description

The data comes from Chengdu Rail Transit Group [6]. It includes a total of 181 antenna network total passenger trips from January 1, 2025, to June 30, 2025, and daily passenger flow trips of 16 lines, totaling 3077 data. Among them, the average value of passenger traffic is 6,027,200, the minimum value is on January 28, which is only 1,814,600, the maximum value occurs on April 30, which is 8,238,000, the 25% quartile is 5,248,900, the median is 6,462,600, and the 75% quartile is 6,738,800 passengers. There are no missing values in the raw data, and no missing value processing is required.

2.2 Smoothness test and first-order difference treatment

Fig. 1 shows a rolling statistical graph; the horizontal coordinate indicates the date, the vertical coordinate is the passenger flow in ten thousand, and the black solid line indicates the original passenger flow data, excluding special holidays. The passenger flow of the Chengdu metro shows obvious cyclical changes. The sequence shows an obvious trough in February 2025, which may be due to the influence of the Spring Festival holiday. For holidays and other outliers, they are

* Corresponding author: scuzyj2005@outlook.com

 https://doi.org/10.1051/itmconf/20268801005

retained because they reflect the real passenger flow characteristics. The red line shows a sharp drop between January and February and stabilizes after rebounding. The green dashed line is the seven-day moving standard deviation, which also shows a significant peak in February, indicating that the mean and variance change over time. Thus, it can be visualized that the original series is not smooth.

The ADF test is commonly used in time series analysis of the smoothness test to determine if there is a unit root. After the calculation, the lagged order used is 13, and the number of observations is 167. It is derived from the ADF test that the p -value is 0.142 > 0.05, the series is non-stationary; the ADF statistic (-2.398) is greater than all the critical values, which indicates that there is a unit root, and a first-order differencing is needed to make the series stationary.

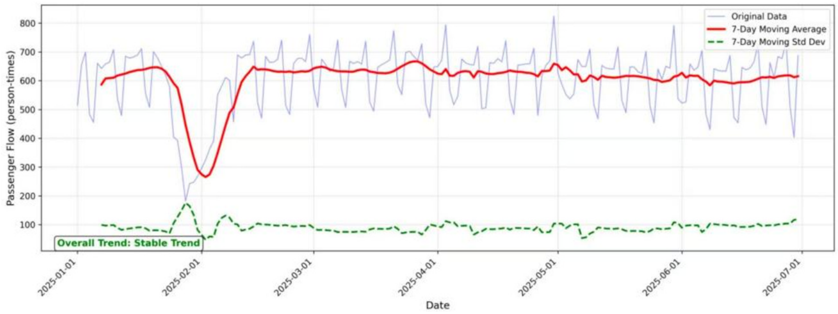

Fig. 1. Rolling Statistical Graph (Picture credit: Original)

Fig. 2 is the 1st order differenced series, with the horizontal coordinates indic ating the time, ranging from January 1, 2025, to June 30, 2025, and the vertical coordinates indicating the amount of change in the original data at the before and after time points. The differenced series fluctuates up and down around the mean value. The ADF test p -value is 0, which is much less than 0.05; the series after 1st order differencing is smooth.

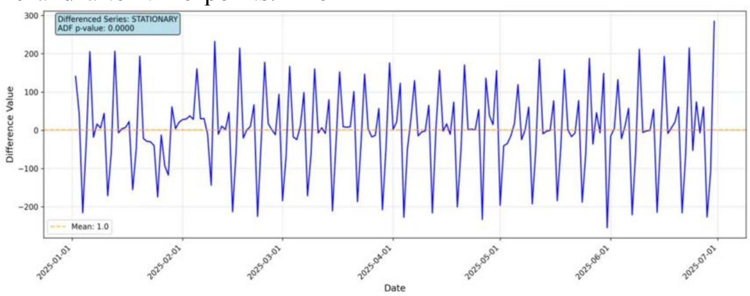

Fig. 2. 1st Order Differenced Series (Picture credit: Original)

2.3 Periodic analysis

In order to study the periodicity of subway passenger flow, Nie found that the top three factors affecting the size of the short-term passenger flow were, in order, the previous time period's passenger flow, the previous week's simultaneous flow, and the previous day's simultaneous flow [7]. Autocorrelation analysis is a measure of the correlation between a time series at different points in time, i.e., the correlation between the series itself and its lagged version.

Fig. 3 shows the autocorrelation function ACF, where the horizontal coordinate indicates the number of lag days and the vertical coordinate indicates the autocorrelation coefficient. The autocorrelation analysis reveals that the autocorrelation coefficients of 1-day lag and 7-day lag are more than 0.5, indicating that today's traffic is highly correlated to yesterday's traffic and there is obvious weekly periodicity. There is significant autocorrelation and weekly seasonality in the original series.

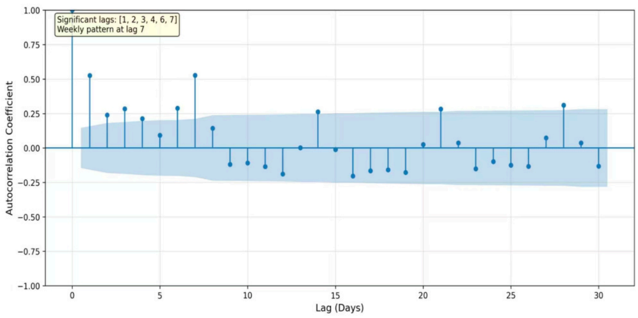

Fig. 3. Autocorrelation Function ACF (Picture credit: Original)

Fig. 4 shows the partial autocorrelation function PACF. The horizontal and vertical coordinates of it are roughly the same as those of the ACF plot. As can be seen from Fig. 4, the partial autocorrelation coefficients of lag 1 and 2 are significantly not 0. From lag 3, the coefficients basically fall into the confidence interval, i.e., the PACF of the original series ends after lag 1, and the AR order p=1 is initially determined.

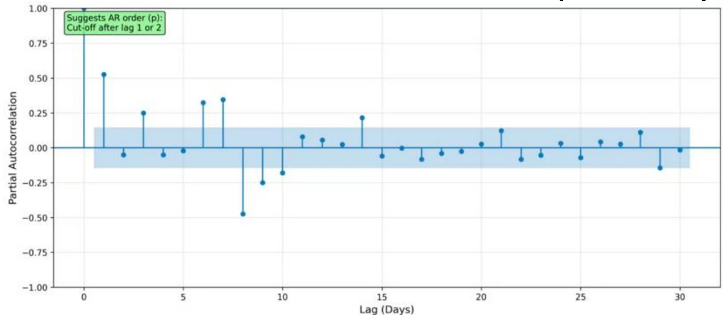

Fig. 4. Partial Autocorrelation Function PACF (Picture credit: Original)

2.4 Model diagnosis

After the time series is constructed to simulate a combined ARIMA model, a residual autocorrelation test is usually required to check whether there is still an unextracted autocorrelation pattern. If the residuals are white noise, the model has adequately extracted the predictable information from the data. After testing, the final result p -value of 0.421 is greater than 0.05, so there is no autocorrelation in the residuals.

3 Results and discussion

After all the above tests, the final model was confirmed as ARIMA(1,1,1)(1,0,1). The researcher used Mean Squared Error (MSE), Root Mean Squared Error(RMSE), Mean Absolute Error(MAE), Mean Absolute Percentage Error(MAPE), and Coefficient of Determination(R^2) to evaluate the model. The calculations resulted in an MSE value of 1910.41, the RMSE is 43.71, the MAE is 31.34, the MAPE is 5.23%, and R^2 is 0.36, which gives a good prediction result that can be used for general business forecasting and can guide decision-making.

Fig. 5 shows the time series plot of historical data and forecast results, the horizontal coordinate is the date, the time range is from the beginning of July 2025 to the end of July 2025, the vertical coordinate is the passenger flow in 10,000 person-times, the blue solid line is the original passenger flow data, and the red dotted line is the forecast data, and the forecast data continues the cyclical fluctuation pattern of the historical data, with the average passenger flow projected to be 6,320,000 person-times per day and fluctuating between 5,020,000 person-times per day to 7,190,000 person-times per day. The average flow is expected to be 6.32 million passengers per day, and fluctuates between 5.02 million passengers per day and 7.19 million passengers per day, with a standard deviation of 69, which is consistent with the pattern of fluctuation in the later period of history. The pink shading is the 95% confidence interval, which ranges from about 1 million to more than 10 million passengers per day, and is wide enough to indicate that there is a 95% probability that future traffic will fall within this interval, and the width of the interval reflects the uncertainty of the forecast.

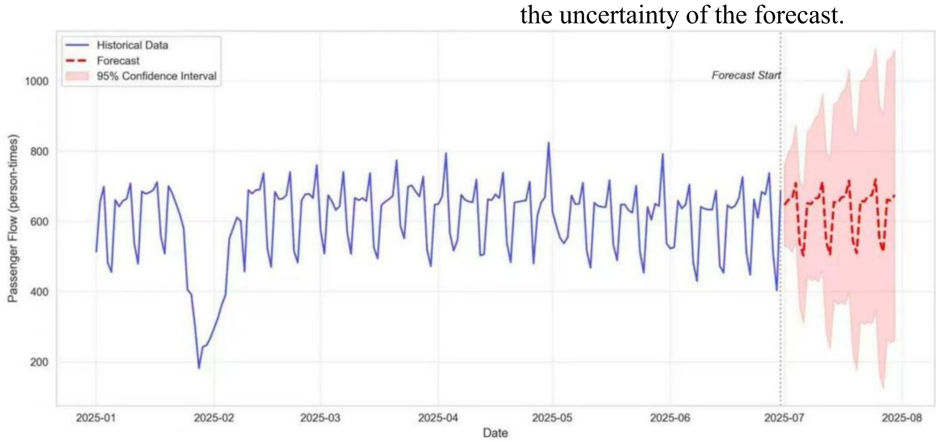

Fig. 5. Time Series Plot of Historical Data and Forecast Results (Picture credit: Original)

4 Discussion

The standard deviation of this forecast result is only 69, which is much lower than the fluctuations in the historical data, indicating that the degree of dispersion of passenger flow in the next 30 days is low, and the overall operation will remain relatively stable. The extreme range from 5.02 to 7.19 million passenger trips, and the mean value of 6.32 million passenger trips, form a clear operating range of passenger flow, covering most of the core fluctuation ranges in the historical data. The wide range characteristic of the 95% confidence interval indicates that while focusing on the core forecast value, the potential risk of extreme cases must be emphasized.

Of course, there are some limitations in this prediction, and there is still room for optimization in the future. Liu proposed that holidays, events, and activities, and short-term fare fluctuations are all factors of metro passenger flow [8]. These factors may lead to a significant deviation of passenger flow from the forecast interval. In the future, the prediction system can be optimized in several directions, such as introducing external data, such as weather, holidays, policy announcements, etc., to enrich the model input variables and improve the responsiveness to unexpected factors.

In the future, more models can be used to improve forecasting accuracy. Zhou established a combination model of the ARIMA model and the neural network model, which improved the prediction accuracy [9]. Lv predicted highway traffic and transportation volume based on the ARIMA-Long Short-Term Memory (LSTM) combination model, and concluded that the combination model improved in prediction accuracy compared with a single model [10]. Yu et al., by establishing two passenger flow prediction models, ARIMA and LSTM, concluded that predicting the same set of target data, the prediction accuracy of the LSTM model is better, and the computational speed of the ARIMA model is faster. In terms of model adaptability, the ARIMA model is better than the LSTM model for predicting relatively long time dimensions [11]. A dynamic updating mechanism can also be established to periodically adjust the prediction model based on the latest data to ensure that the prediction results fit the actual situation.

5 Conclusion

Overall, this forecast quantifies the change of passenger flow from July 1 to July 30, 2025, through the time-series ARIMA model based on the data in the first half of 2025. The average value of the 30-day forecast passenger flow is 6.32 million passengers/day, the maximum value can reach 7.19 million passengers/day, the minimum value is 5.02 million passengers/day, and the standard deviation is 69. This result shows that the core fluctuation range of the future passenger flow is concentrated in the range of 5 to 7.2 million passengers, and the daily average level is basically the same as the historical average value from March to June 2025.

It also clarifies the core trend and the fluctuation range of future passenger flow, which provides important data support for operation management, but still needs to pay attention to the potential risks. In the future, with the richness of data accumulation and the iteration of model technology, the passenger flow forecast will be more accurate, becoming a core tool to promote intelligent and refined operation management, and injecting new impetus for the industry's high-quality development.

References

1. W. Cheng, Rail transit "fourth city"--Chengdu metro enters the era of large line network. Transp. Constr. Manage. (06), 124-126 (2020)
2. D. XI, Study on the Influencing Factors of Mass Passenger Flow in Metro Stations and Optimization of Operation Organization. Transp. Manage. World (32), 1-3 (2025)
3. X. Zhu, Research on Price Forecasting of Small Agricultural Products in China Based on ARIMA Model--Taking the Market Price of Chili Pepper as an Example. Price Theory Pract. (12), 226-230+295 (2025)
4. M. Liao, On the practical application of ARIMA mathematical model. Proc. Int. Forum Acad. Res. Educ. Sci. Dev. (IX) (2025)
5. Q. Wang, Characteristics of spatial and temporal distribution of passenger flow and ARIMA passenger flow prediction in Wuhan-Hamzhou Intercity Railway. Railw. Oper. Technol. **32**(01), 22-26 (2026)
6. Chengdu Rail Transit Group, Daily passenger flow. (2025) www.chengdurail.com
7. X. Nie, D. Wu, J. Jing, et al., Analysis of short-time inbound passenger flow prediction and influencing factors of subway based on LightGBM and SHAP. Railw. Econ. Res. (05), 50-55 (2023)
8. Y. Liu, Analysis of the impact of sensitivity factors on urban rail transit passenger traffic. Transp. Eng. **17**(03), 14-18 (2017)
9. Y. Zhou, L. Zhou, Y. Le, Research on short-time passenger flow prediction of urban rail transit based on combined model. Proc. Int. Conf. Eng. Bus. Manage. (EBM2010) (2010)
10. Z. Lv, P. Wang, Highway traffic volume prediction based on ARIMA-LSTM. Proc. 42nd China Control Conf. (15) (2023)
11. Y. C. Yu, S. Li, Z. Liu, Comparison of passenger flow prediction methods at railroad stations based on ARIMA and LSTM. J. Shanghai Univ. Eng. Technol. **38**(03), 278-283 (2024)