Medium-Term Forecast of Metro based on Big Traffic Data

Xiaoyu Ma*

Department of Transportation Engineering, Shandong Jianzhu University, Jinan, China

* Corresponding Author Email: 1910841216@mail.sit.edu.cn

Abstract. With the rise of industrialization and automation, the demand for transportation has been continuously increasing. The demands for quick and effective travel can no longer be met by conventional transportation techniques. In the field of transportation, urban train travel is significant. At the same time, passenger flow prediction has become increasingly important. In addition to short-term forecasts, medium-term forecasts of daily intervals of passenger flow are also essential. This research employs the ARIMA model to forecast medium-term passenger flow, utilizing subway passenger flow data collected in Chengdu City from April 29th to September 28th, 2019. Based on the data and methodology, this study investigates the feasibility of using the ARIMA model for medium-term subway passenger flow prediction. The research findings indicate that the model exhibits strong seasonality, with an ADF unit root test p-value less than 0.05 and an LB statistic test p-value less than 0.05. The Mean Percentage Error is only -1.03%. Furthermore, there is a strong concordance between the fluctuation patterns observed in the actual and predicted values of the model, which closely align with the trends observed in the training dataset. Through this research, it is concluded that the ARIMA model performs well in medium-term subway passenger flow prediction.

Keywords: Metro, ARIMA, SARIMA, passenger flow forecasting.

1. Introduction

In light of the burgeoning industrialization and technological advancements, there has been a notable surge in the populace's transportation requirements, rendering traditional modes of transportation inadequate in meeting the escalating need for swift and efficient travel. With the invention and popularization of cars, trains, airplanes, and other means of transportation, transportation has been greatly improved. At the same time, the development of information technology has also provided new solutions for transportation, such as intelligent transportation systems, shared trips, and so on. In the realm of transportation, urban train travel is crucial. Passenger flow forecasting has always been an important aspect of relevant research and practice when building and operating metropolitan rail transportation, and accurately predicting the number of future passenger arrivals is essential to ease operational pressure. Despite the successful implementation of intelligent video analysis technology for real-time monitoring of passenger numbers in railway passenger stations, accurately predicting future passenger flow trends remains a formidable challenge.

During the practical forecasting procedure for metropolitan rail transportation, the temporal span of the anticipated data typically ranges from 15 to 60 minutes. Nevertheless, apart from the short-term prognostication of rail transit passenger volume, the medium-term estimation of passenger flow utilizing daily temporal intervals assumes greater significance. A relatively accurate analysis of the long turnover of people is not possible in short-term forecasts. Longer forecast time intervals (day as the forecast time interval) will be more common in the medium-term forecast adaptation.

At present, the relevant forecasting methods usually include time series forecasting, grey forecasting, and neural network forecasting, among which $ARIMA$ forecasting model and $LSTM$ forecasting model are the most efficient and fast. Autoregressive Moving Average Model ($ARIMA$) is a commonly used time series analysis method, which is used to predict future data points. It combines the concepts of autoregression (AR) and moving average (MA) to model the trend, seasonality, and randomness of time series data [1-3]. Zhang used $ARIMA$ model to predict the changes of bearing vibration signals in the short term in the future, and input the prediction results into $XGBoost$ model for fault classification prediction, so as to realize fault identification of rolling bearings and improve the prediction accuracy [4]. To effectively capture the attributes of time series

data pertaining to wind power generation, Tu et al. put forward a novel technique for modeling the marginal distribution, specifically known as the autoregressive comprehensive moving average-generalized autoregressive conditional square moment-T ($ARMI-Garcht$) model [5]. Based on the QoS prediction model, Yan extended the $ARIMA$ model so that it can effectively predict multiple QoS values at the same time [6].

The Long Short-Term Memory ($LSTM$) network is a specific type of recurrent neural network (RNN) that incorporates time loops, specifically engineered to address the challenge of long-term dependencies encountered by conventional RNNs. Due to its unique design structure, $LSTM$ is highly appropriate for the processing and prognostication of pivotal occurrences within time series exhibiting exceedingly protracted intervals and delays. As an intricate nonlinear component, $LSTM$ can serve as a sophisticated nonlinear module for the construction of more extensive deep neural networks [7]. In addition, $LSTM$ can identify the structure and pattern of data, can mine the nonlinearity and complexity contained in data, and is widely used in the prediction research based on time series [8,9]. Li combined the improved Particle Algorithm ($IPSO$) with the $LSTM$ model, and proposed the $IPSO-LSTM$ rail passenger volume prediction model [10].

By constructing $ARIMA$ model, this paper analyzes and proves the actual data of $Gitong$ passenger flow on urban rail to verify the correctness and accuracy of the model. At the same time, in the case that $LSTM$ model is mostly used in short-term passenger flow prediction, on the basis of the former, by building relevant models and analyzing the data, it is tested whether $LSTM$ model can be used in the medium-term forecast of urban rail transit passenger flow.

2. Methods

2.1. Data Source

This study utilizes the objective and accurate data source of passenger flow data from Chengdu Line 1, spanning from April 29 to August 28, 2019.

2.2. Indicator Selection and Description

Table 1 shows the full names, data types, and explanations of the three variables used in the study. Counting the date can record the information of passengers entering and leaving the station well, and the accuracy rate is high, which makes the data of the study more authoritative. The line number is the name of the line that is targeted, which is highly targeted.

Table 1. Name and explanation of variables

Full Name	Data type	Explanations
Date of statistics	Typedef	Inbound time
Line number	INT	Ride line number
Number of stops	INT	Foot traffic on the line

2.3. Method Introduction

2.3.1 ARIMA model

$ARIMA$ model, differential integrated moving average the model, also known as integrated moving average the model (moving can also be called sliding), is one of the predictive analysis methods of time series. In $ARIMA(p,d,q)$, AR is "autoregressive", and the variable p represents the count of autoregressive components, while MA denotes the "moving average" aspect, with q signifying the number of moving average terms. Additionally, d corresponds to the number of differencing operations performed to achieve stationarity within the sequence.

After eliminating the local level or trend component from a non-stationary time series, it exhibits a certain degree of homogeneity, indicating that certain segments of the series bear resemblance to other segments. By subjecting this type of non-stationary time series to differencing, it can be

transformed into a stationary time series. Such a time series is referred to as a homogeneous non-stationary time series, with the order of differencing representing the degree of homogeneity.

If it is denoted as the difference operator, then there is:

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

Upon receipt of the manuscript, it is presumed that the corresponding authors have conferred upon us the copyright authorization to utilize the paper within the context of the relevant book or journal. Upon receipt of the manuscript, it is presumed that the corresponding authors have granted us the copyright permission to utilize the paper.

For the delay operator B, the paraphrase is:

$$y_{t-p} = B^p y_t \,, \forall p \geq 1 \tag{2}$$

It follows that:

$$\nabla^k = (1 - B)^k \tag{3}$$

With a homogeneous nonstationary sequence y_t of order d, then if there is a $\nabla^d y_t$ is characterized as a time series exhibiting stationarity, it can be set to an $ARMA(p, q)$ model:

$$\lambda(B)(\nabla^d y_t) = \theta(B)\varepsilon_t \tag{4}$$

Among them, $(B) = 1 - \lambda_1 B - \lambda_2 B^2 - \cdots - \lambda_p B^p$ and $\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_p B^p$ are autoregressive coefficient polynomial and the moving average coefficient polynomial, respectively. ε_t is zero mean white noise sequence. The autoregressive integrated moving average model, denoted as $ARIMA(p, d, q)$, can be referred to as the suggested model. In the case where the differencing order d is zero, the $ARIMA$ model is essentially identical to the autoregressive moving average model ($ARMA$). The key distinction between these two models lies in the presence or absence of differencing, specifically whether the differencing order d is zero, indicating the stationarity of the time series.

2.3.2 SARIMA Model

The quarterly differential autoregressive moving equilibrium model ($SARIMA$ model) is an improvement on the ($ARIMA$ model), which is used to transform non-smooth time series into smooth periodic series. The model achieves the difference and autoregressive moving average of the time series by regression of the current and lagging values of the dependent variable, as well as taking into account the random error term. This transformation makes the time series smoother and better able to capture periodic features. The $SARIMA$ model is an analytical method suitable for medium-length or even longer time series data. It can not only deal with irregular and non-smooth time series data, but also take into account common seasonal features in time series. In the $SARIMA$ model, the original periodic sequence y_t is first differentiated periodically to eliminate tendency. Quarterly differentiation is then done to remove seasonality. This differential operation smooths out the time series while retaining important cyclical features. After the above processing process, the resultant model can be mathematically represented as follows:

$$\mathrm{SARIMA(p, d, q)} * \mathrm{(P, D, Q, S)} \tag{5}$$

Be denoted as:

$$\varphi p(B)\Phi P(Bs)(1 - B)d(1 - Bs)Dy_t = \theta q(B)\Theta Q(Bs)u_t \tag{6}$$

Or it can be written as follows.

$$\varphi p(B)\Phi P(Bs)\nabla d\nabla sDy_t = \theta q(B)\Theta Q(Bs)u_t \tag{7}$$

where $(1 - B)D$ and∇D both represent difference operators; $(1 - Bs)D$ and $\nabla s\, d$ represent the seasonal difference operator; d represents the phase-by-period differential order; S stands for phase difference step length; u_t Stands for white noise sequence; $\varphi p(B)$ Represents the non-

quarterly self-return formula, $\varphi p(Bs)$ represents the quarterly P-order self-return operator polynomial, P represents the autoregressive order, p represents the quarterly autoregressive order; $\theta Q(B)$ Represents the expression of non-quarterly moving equilibrium, $\theta Q(Bs)$ represents the expression of quarterly Q order moving equilibrium operator, q represents the order of moving equilibrium.

3. Results and Discussion

3.1. Descriptive Analysis

This study utilizes the time period from April 29, 2019, to August 31, 2019, as the training set, while the time period from September 1, 2019, to September 28, 2019, is designated as the test set. Initially, the training set is visualized using the R programming language, and the sequential plot of the complete training set is depicted in Figure 1.

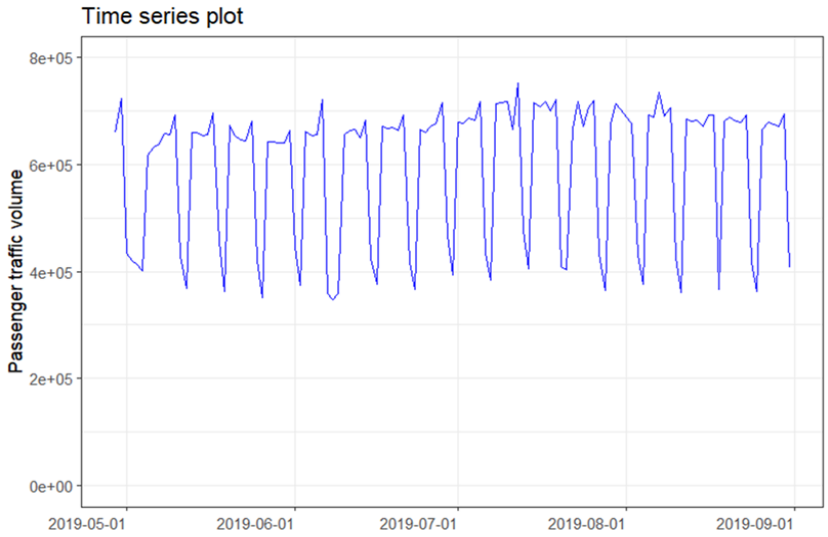


Fig. 1 Sequence diagram of training set

Periodicity can be preliminarily found by observing the figure above, and it is sequentially decomposed, as shown in Figure 2.

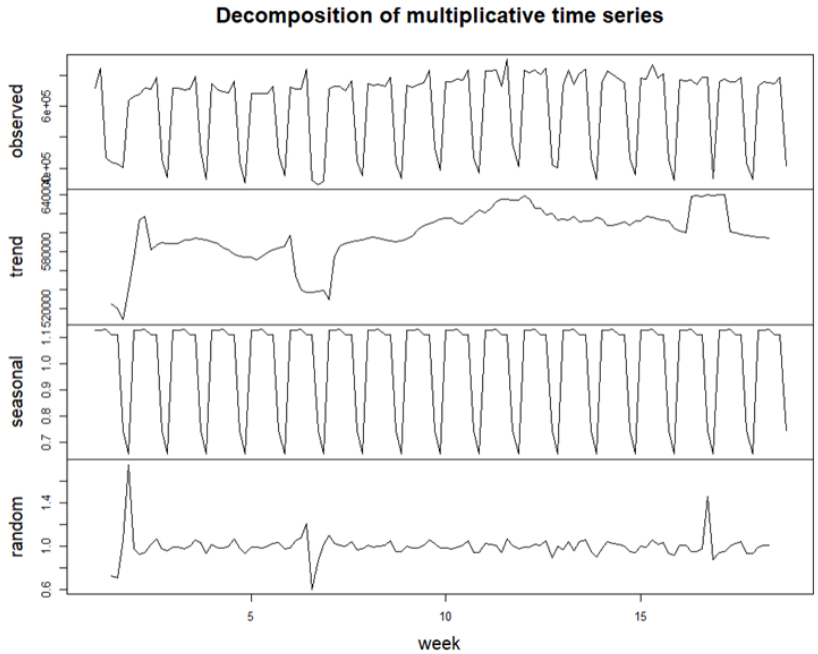


Fig. 2 Time series decomposition diagram

In the diagram, the initial stage represents the primary temporal data series, while the subsequent stage pertains to the behavioral data trend, denoting a persistent and evolving pattern or condition observed over an extended duration. The third is the seasonal factor, which is the regular change of the development level of the phenomenon caused by the change of seasons; The last one is random change, which refers to the influence of many accidental factors on the time series. Observe that the decomposed seasonal factors in the figure above have obvious regular changes, indicating seasonality. Therefore, the SARIMA model is used for prediction. The original data were processed by the first 7-step seasonal process, and then ACF and PACF plots were drawn. As shown in Figure 3 and 4.

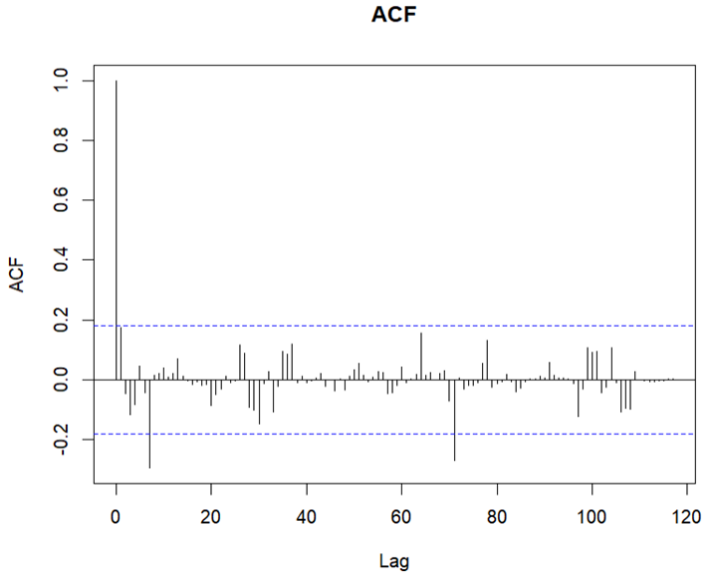


Fig. 3 ACF diagram

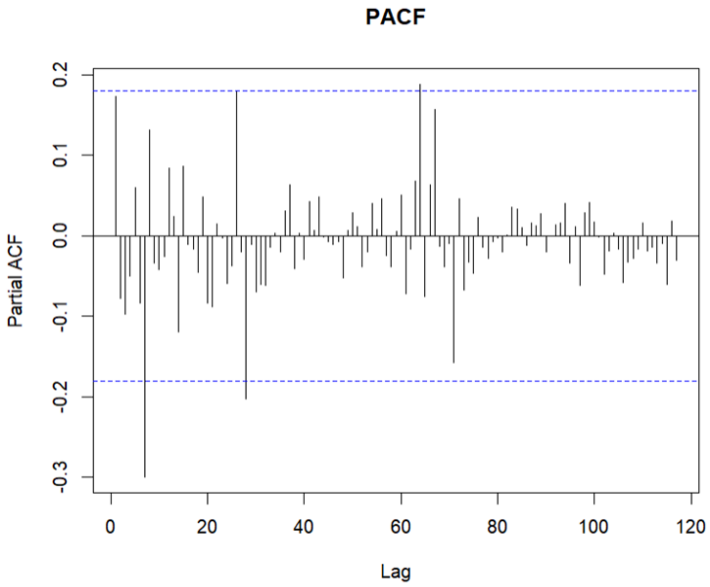


Fig. 4 PACF diagram

For the sake of objectivity, various statistical tests for sequence stationarity have been applied, the most widely used of which is the unit root test. If the sequence is smooth, there is no root of identity, and the stationary nature of the sequence can be tested by constructing an ADF test statistic. When the null hypothesis H_0, at least one unit root exists in the sequence; When optionally assuming H_1, there is no root of identity for the sequence. When the t-statistic in the root of unity test result is much less than the significance level α, and the p-value representing the rejection of the null hypothesis for the first type of error is close to 0, it can be considered that there is no root of the unit, that is, the sequence is stationary.

Pass the Augmented Dickey-Fuller Test, the p-value is equal to 0.01 and less than 0.05, there is no unit root, indicating that the time series of 7-step seasonal difference is stable.

In a time, only when the historical data of the series has a certain impact on future development, such a series is researchable and it is worth mining and predicting future series. If the values within the sequence exhibit no discernible interdependence, meaning that past occurrences have no influence on present or future developments, the sequence is classified as a purely stochastic sequence. The pure randomness test, also known as the white noise test, is tested for pure randomness by constructing test statistics on the basis that the sequence has been stable, and in order to determine its validity.

The Barlett Proof method shows that in the case of a time series being purely stochastic, yielding an observation sequence comprising n periods, the sample autocorrelation coefficient for a non-zero lag of the sequence approximately adheres to a mean of 0, with a variance following a normal distribution inversely proportional to the observation period of the sequence.

$$\hat{\rho}_k \sim N\left(0, \frac{1}{n}\right), \forall k \neq 0 \tag{8}$$

The null hypothesis H_0: Sequence independence with delay periods less than or equal to m.

$$H_0: \rho_1 = \rho_2 = \cdots = \rho_m = 0 \tag{9}$$

Alternative hypothesis H_1: Sequences with a lag period less than or equal to m periods are correlated.

$$H_1: \rho_1, \rho_2, \ldots, \rho_m \text{not all equal to } 0 \tag{10}$$

Pass the Augmented Dickey-Fuller Test, the LB statistic test p-value is equivalent to 0.01164. For values below 0.05, the sequence can be classified as a non-white noise sequence, and non-pure random is valuable.

The seasonal $ARIMA(p,d,q) \times (P,D,Q)_S$ process is as follows.

$$\phi(B)\Phi(B^S)Y_t = \theta(B)\Theta(B^S)Z_t,\ Z_t \sim WN(0,\sigma^2) \tag{11}$$

$$\phi(x) = 1 - \phi_1 x - \cdots - \phi_p x^p, \Phi(x) = 1 - \Phi_1 x - \cdots - \Phi_P x^P \tag{12}$$

$$\theta(x) = 1 + \theta_1 x + \cdots + \theta_q x^q, \Theta(x) = 1 + \Theta_1 x + \cdots + \Theta_Q x^Q \tag{13}$$

Forecast package for the R programming Language, it provides the auto.arima() function and automatically determines the order based on the AIC criterion. By using the aforementioned process and function, the identified model is $ARIMA(0,1,1) \times (0,1,1)_7$, namely:

$$\phi(x) = 1 \tag{14}$$

$$\Phi(x) = 1 \tag{15}$$

$$\theta(x) = 1 + \theta_1 x \tag{16}$$

$$\Theta(x) = 1 + \Theta_1 x \tag{17}$$

$$S = 7 \tag{18}$$

$$Y_t = (1 + \theta_1 B)(1 + \Theta_1 B^7)Z_t, Z_t \sim WN(0,\sigma^2) \tag{19}$$

Expanding the above equation yields the following results:

$$Y_t = (1 + \Theta_1 \nabla_7 + \theta_1 \nabla + \theta_1 \Theta_1 \nabla_8) Z_t \tag{20}$$

Substitute the parameters $\theta_1 = 0.2385, \Theta_1 = -0.8562$ into the above equation:

$$Y_t = Z_t + 1.2385 Z_{t-1} - 0.8562\, Z_{t-7} - 0.2043\, Z_{t-8}, Z_t \sim WN(0,\sigma^2) \tag{21}$$

Table 2 shows the output training set error parameters.

Table 2. Prediction result error

-	MPE	MAPE	MASE	ACF1
Training set	1.03493	6.976351	1.026656	0.001210975

It can be seen that the average error percentage is small, and the average percentage error MPE is only -1.03%, and the equation performs well on the training set.

3.2. Inference Analysis

In order to visually demonstrate the fitted and predicted values of the aforementioned model, data visualization was performed on the existing model and fitted data. As shown in Figures 5 and 6, the predicted results of the test set are represented by the blue line, and the fluctuation trend is close to that of the training set, indicating a good fit of the model.

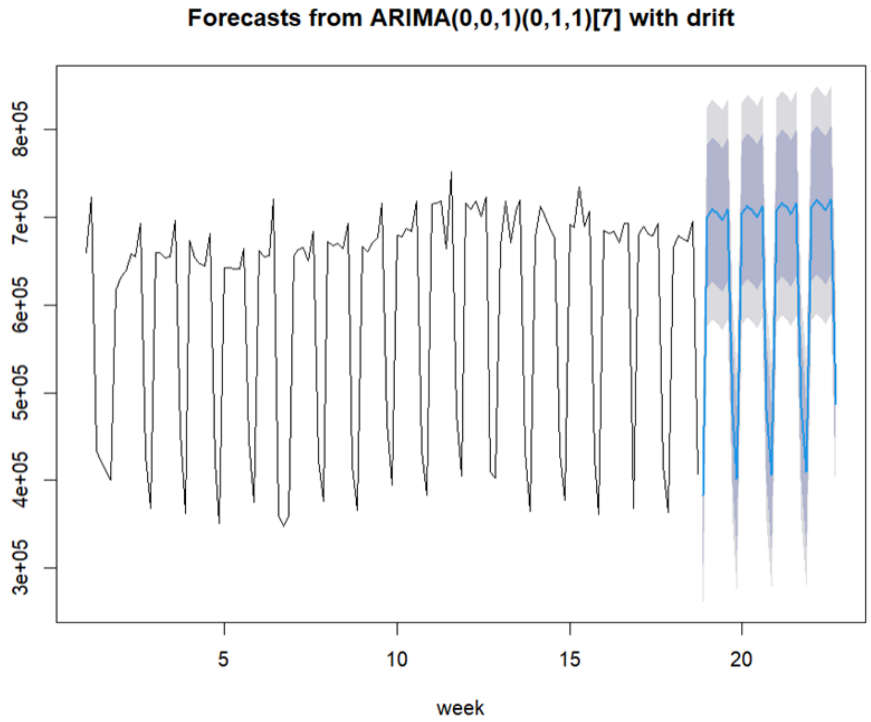


Fig. 5 Prediction graph

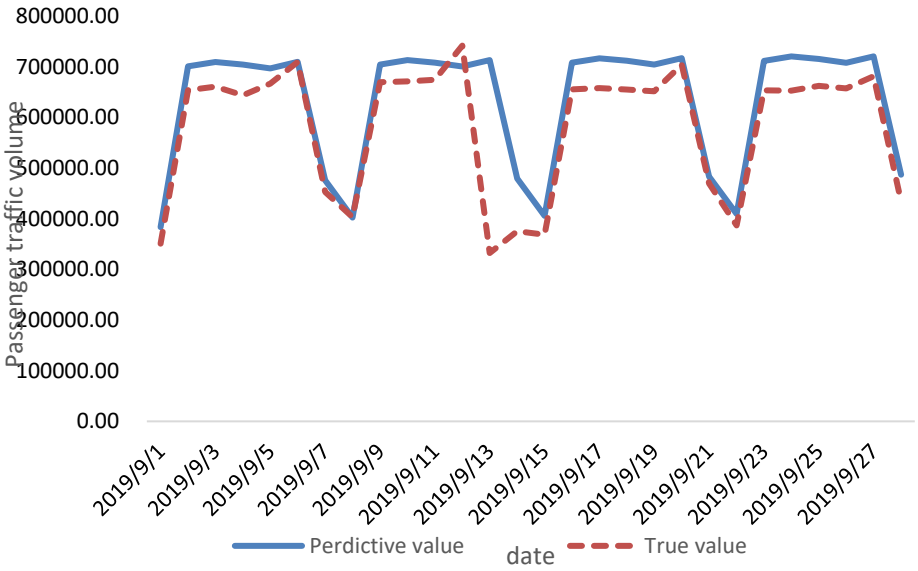


Fig. 6 The predicted value compared with the true value

A common approach to evaluating the adequacy of a statistical model constructed using a dataset involves comparing the observed values with the corresponding predicted values derived from the fitted model. When the fitted model is suitable, the residuals should exhibit a certain degree of conformity with the model. The residuals were subjected to a white noise test, and the Ljung-Box test was used to obtain a chi-square value of 0.095945 with 1 degree of freedom. The obtained p-value is 0.7568, the hypothesis that the residuals are white noise cannot be rejected. Therefore, the test is considered to be successful. The predicted results and the comparison and error with the real data are shown in Table 3.

Table 3. Prediction result error

Time	MPE	MAPE	MASE
2019/9/1	350569	383285.5	9.33%
2019/9/2	653608	700444	7.17%
2019/9/3	660333	709124.3	7.39%
2019/9/4	643466	704188.3	9.44%
2019/9/5	666520	696600.6	4.51%
2019/9/6	709116	709269.8	0.02%
2019/9/7	452774	475677.9	5.06%
2019/9/8	403010	402004.5	- 0.25%
2019/9/9	669414	704098.3	5.18%
2019/9/10	670783	712778.7	6.26%
2019/9/11	674097	707842.7	5.01%
2019/9/12	741492	700254.9	- 5.56%
2019/9/13	332147	712924.1	114.64%
2019/9/14	375394	479332.2	27.69%
2019/9/15	368281	405658.8	10.15%
2019/9/16	655023	707752.7	8.05%
2019/9/17	657621	716433	8.94%
2019/9/18	654749	711497	8.67%
2019/9/19	651313	703909.2	8.08%
2019/9/20	703584	716578.4	1.85%
2019/9/21	469530	482986.5	2.87%
2019/9/22	386552	409313.1	5.89%
2019/9/23	653276	711407	8.90%
2019/9/24	652658	720087.3	10.33%
2019/9/25	661914	715151.3	8.04%
2019/9/26	657139	707563.5	7.67%
2019/9/27	681080	720232.8	5.75%
2019/9/28	433636	486640.8	12.22%

Through the above research and exploration, the passenger flow of the subway station is intuitively reflected. The findings align with the evolving pattern of passenger flow, demonstrating a close resemblance between the projected trend and the observed trend. Simultaneously, the changing trend of passenger flow on different dates is consistent with the residents' travel situation, and the error is relatively small. Finally, the forecast effect is good, which further validates the feasibility of $ARIMA$'s medium-term forecast of passenger flow.

4. Conclusion

This paper aims to explore the feasibility and effectiveness of using the Autoregressive Integrated Moving Average ($ARIMA$) model to predict subway passenger flow. By analyzing and modeling subway passenger flow data, it was found that the $ARIMA$ model has certain advantages and potential applications in predicting subway passenger flow.

Firstly, this paper collected subway passenger flow data over a period of time and conducted preprocessing and analysis. By observing the trends, seasonality, and randomness of the data, the suitability of the $ARIMA$ model was determined. Then, based on the autocorrelation and partial autocorrelation functions of the data, appropriate $ARIMA$ model orders were selected, and the selected $ARIMA$ model was used to forecast subway passenger flow. Through a comparative analysis of the projected outcomes and the observed data, it was determined that the ARIMA model exhibits a high degree of efficacy in accurately capturing the dynamic patterns of subway passenger flow. The accuracy and stability of the forecasted results indicate a certain reliability of the $ARIMA$ model in predicting subway passenger flow.

However, this paper also identified some limitations of the $ARIMA$ model. Firstly, the $ARIMA$ model assumes that the data is stationary, but actual subway passenger flow data may exhibit non-stationarity. Therefore, before applying the $ARIMA$ model, it is necessary to test and transform the data for stationarity. Secondly, the $ARIMA$ model is sensitive to outliers and extreme values, which may lead to deviations in the forecasted results. Therefore, during the modeling process, it is important to handle outliers or consider using other models to address exceptional cases.

In conclusion, the $ARIMA$ model has certain advantage and potential applications in predicting subway passenger flow. However, it is important to be aware of the limitations of the $ARIMA$ model and consider incorporating other methods and techniques to improve the accuracy and stability of the prediction. Future research can explore other time series models or combine machine learning algorithms to enhance the prediction of subway passenger flow.

References

[1] Zhang Li. Research and system implementation of analysis and prediction algorithm based on time series ARIMA model. Jiangsu University, 2008.

[2] Hou L. Short-term analysis and prediction of oil price based on ARIMA model. Jinan University, 2009.

[3] Xia L. Research on forecasting method of regional electricity consumption based on ARIMA model and regression analysis. Nanjing University of Science and Technology, 2013.

[4] Zhang Tianrui, et al. Research on Rolling Bearing Fault Prediction Model Based on ARIMA and XGBoost. Manufacturing technology and machine tool, 2022 (4): 176-182.

[5] Qingyu Tu, Shihong Miao, Fuxing Yao, et al. Forecasting Scenario Generation for Multiple Wind Farms Considering Time-series Characteristics and Spatial-temporal Correlation. Journal of Modern Power Systems and Clean Energy, 2021, 9(4): 837-848.

[6] Yan Chao, Zhang Yankun, Zhong Weiyi, et al. A Truncated SVD-Based ARIMA Model for Multiple QoS Prediction in Mobile Edge Computing. Tsinghua Science and Technology, 2022, 27(2): 315-324.

[7] TIAN Y, PAN L. Predicting Short-term Traffic Flow by Long Short-term Memory Recurrent Neural Network. IEEE International Conference on Smart City. Chengdu: IEEE, 2015, 153-158.

[8] Ren Na. Short-term Passenger flow prediction of urban rail transit based on deep Learning Algorithm. System simulation technology, 2021, (4): 259-264.

[9] CUI Hongtao, Chen Xiaoxu, Yang Chao et al. Metro Inbound passenger flow prediction based on deep long Short-Term Memory Network. Urban Rail Transit Research, 2019, 22(09): 41-45.

[10] Li Wan, Feng Fenling, Jiang Qiwei. Improved particle swarm algorithm to optimize the LSTM neural network rail passenger traffic prediction. Journal of railway science and engineering, 2018, (12): 3274-3280.