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

Research on the Metro Ridership Forecasting based on ARIMA Model

ShiRu Lyu*

Faculty of Data Science, City University of Macau, Macau, 999078, China

*Corresponding author: D23090105821@cityu.edu.mo

Abstract. With the increasing coverage of subways in cities, people travel more than just by bus or walking. Nowadays, subway stations have become sites with high population density in the city. Due to the increasing travel demand of residents. Excessive congestion in subway stations will lead to inconvenience and even accidents. The subway passenger flow prediction can avoid many potential problems. Passenger flow prediction can not only optimize the subway scheduling, but also help the operating company to fully prepare for the security work during the rush hours. Data from February 1,2023 to December 31,2023 are from Xi'an Transportation Platform. This research used the ARIMA model for mid-term passenger flow prediction, and the data from February to October were used as the training set. The first order of the data is differential to ensure the stability of the data. The values of p and q in the model were determined using the autocorrelation function, and the most appropriate combinations of p and q were selected by BIC. The ARIMA model was built using (p, d, q), and predicted passenger traffic in November and December. Finally, the feasibility of predicting the medium-term subway passenger flow is determined by comparing the predicted value with the true value.

Keywords: Metro; ARIMA; BIC; passenger flow forecasting.

1. Introduction

Nowadays, with the rapid development of information technology and data processing technology, researchers have more accurate control of data. The traditional statistical and prediction methods are gradually becoming obsolete. At the same time, with the rise of computers a lot of artificial algorithms were eliminated. Computers serve as computing and prediction work. In the field of transportation, residents' demand for transportation is extremely increasing. Traffic forecast has been the focus of the researchers' research object to avoid overcrowding and relieve pressure on operating companies. Although with the help of the computer, traffic prediction technology constantly breaks through the bottleneck. Nevertheless, forecasting accurate data remains a challenge. In the study of subway passenger flow, the research focus is mainly on the short-term passenger flow forecast between the time span of 10 minutes and 30 minutes, but the medium-and long-term subway passenger flow forecast in days is equally meaningful. Establishing a scientific and reasonable mathematical model to predict the future subway passenger flow can make the subway operation department more reasonably dispatch vehicles and prepare for emergency measures.

Nowadays, the main methods for predicting passenger flow include time series prediction, grey prediction and neural network prediction. Chang et al. used Back Propagation Neural Network (BPNN) to predict rail transit passenger flow and constantly adjust the weights and thresholds of the network through back-propagation to minimize the sum of squared errors of the network [1]. Wen and Luo established a short-term passenger flow prediction model of bidirectional long-and short-term memory network based on the theoretical framework of deep learning. And compared the prediction results with the prediction results of decision tree model, support vector machine algorithm and long-and short-term memory network [2]. Liu et al. put forward a station entry and exit prediction model based on the prediction problem of subway station passenger flow. The results showed that the model achieved good prediction performance in both short-term and long-term prediction [3]. Li and Wu proposed the subway station passenger flow statistics and the subway station passenger flow

analysis model based on XG Boost and support vector regression machine and used the cross-validation method to solve the problem of high deviation or high variance in the model training process to improve the model performance [4].

However, the most convenient and fastest model in the model is Autoregressive Moving Average Model (ARIMA). Tian et al. cleaned, sliced and normalized the data. Through comparing experiments, Tian found that the Attention Mechanism of Spatio-temporal Long Short-Term Memory Network (ASTLSTM) has lower Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). Relatively high prediction accuracy in multi-step prediction compared with Long Short-Term Memory (LSTM) and Convolutional LSTM Network (Conv LSTM) [5]. Lu et al. used ARIMA to make short-term prediction of Tianjin Line 1 and reached the conclusion that ARIMA could not predict completely and accurately [6]. Wu et al. used the model to predict and analyze the subway passenger flow. Comparing the prediction results with the actual results and the error is controlled within 10 percent [7]. Shao et al. compared the ARIMA with the Seasonal Autoregressive Integrated Moving Average Model (SARIMA) prediction results. Finally, the Support Vector Machine (SVM) and SARIMA were combined to predict passenger flow with the combined model [8]. In the research, Cui et al. integrated the subway IC card data, holiday data and meteorological data making the research can further restore the real situation [9]. Peng et al. found that deep learning model commonly used loss function is difficult to accurately forecast traffic peak. Therefore, introduce the weighted square error for short traffic flow prediction. Peng assigned different weights to the prediction error according to the size of traffic and increased the penalty for the error at the peak of traffic making the neural network paid more attention to the prediction and error at the peak during back-propagation [10]. Li et al. specifically studied the passenger flow characteristics and prediction problems of rail transit under the influence scope of major exhibitions and focused on the outbound passenger flow prediction of stations to support the passenger flow prediction of such special stations around large exhibition stations [11].

2. Methods

2.1. Data Source

This study utilizes the objective and accurate data source of passenger flow data from Xi'an, spanning from February 1 to November 30, 2023. As table 1 shows, the data is stored in xlsx files including Date of statistics and Line number.

2.2. Indicator Description

Table 1 shows the full names, data types, and explanations of the three variables used in the study. Among variables, the date data can well express the time of passengers, and the passenger flow is the container of the number of passengers.

Table 1. Name and explanation of variables

Full Name	Data type	Explanations
Date of statistics	Typedef	Inbound time
Line number	INT	Ride line number
Passenger flow	Float	Number of passengers

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. After the data were stationary. The values of p and q were found by determining the appropriate order of the autoregressive and moving regression processes.

The highest lag point of partial autocorrelation function (PACF) plot can be used to calculate the p value. The highest lag point of the autocorrelation function (ACF) plot can be used to estimate the value of q.

$$\mathrm{PACF(k)} = \frac{\mathrm{E(Z_t - EZ_t)(Z_{t-k} - EZ_{t-k})}}{\sqrt{\mathrm{E(Z_t - EZ_t)^2}}\sqrt{\mathrm{E(Z_{t-k} - EZ_{t-k})^2}}} = \frac{\mathrm{cov[(Z_t - \overline{Z_t}),(Z_{t-k} - \overline{Z_{t-k}})]}}{\sqrt{\mathrm{var(Z_t - \overline{Z_t})}}\sqrt{\mathrm{var(Z_{t-k} - \overline{Z_{t-k}})}}} \tag{5}$$

$$\mathrm{ACF(k)} = \sum_{t=k+1}^{n} \frac{(Z_t - \overline{Z_t}),(Z_{t-k} - \overline{Z})}{\sum_{t=1}^{n}(Z_{t-k} - \overline{Z})^2} \quad (6)$$

After obtaining the three values, the ARIMA model can be fitted and the subsequent results can be predicted. The accuracy of the forecast model can be evaluated when its prediction is compared to the test data.

2.3.2. BIC model

Bayesian information criterion (BIC) is the method used in statistics to select the best model in a finite set of models. This method computes the probability function and adds a penalty term to the number of parameters in the model. This helps to avoid overfitting and provides a balanced approach for model selection. BIC is like the Akaike information criterion (AIC). Both BIC and AIC are widely used for model selection. The penalty term of BIC is larger than AIC. The BIC is an invaluable tool for selecting the best model among the limited options. The penalty term of the BIC for the number of parameters helps to avoid overfitting. Before using the BIC to compare the estimated models when the values of the dependent variable in all the compared estimates are the same. The BIC has the following relationships. (K is the number of model parameters; n is the number of samples and L is the likelihood function).

$$BIC = kln(n) - 2ln(L) \quad (7)$$

The right p and q can be found easily and more convenient through BIC. After obtaining appropriate p and q the ARIMA model can be built, and the subsequent results predicted. Comparing the prediction model with the test data can evaluate the accuracy of the prediction model.

3. Results and Discussion

3.1. Time Series Plot

This study utilizes the time period from February 1, 2023, to October 31, 2023, as the training set, while the time period from November 1, 2023, to December 31, 2023, is designated as the test set. Initially, the programming language is used to visualize the data and then the original data is less stable through observation. The data visualization diagram is shown in Figure 1.

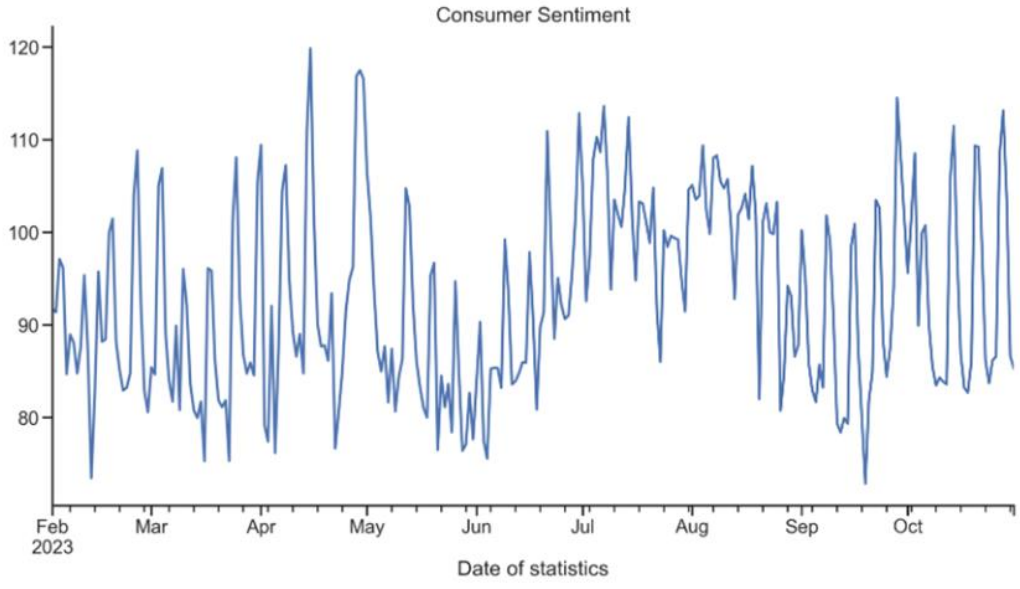

Fig. 1 Sequence diagram of training set

On account of the model limits the stationarity of the data, in order to obtain the first-order difference of the stationary data, the difference plot is shown in Figure 2.

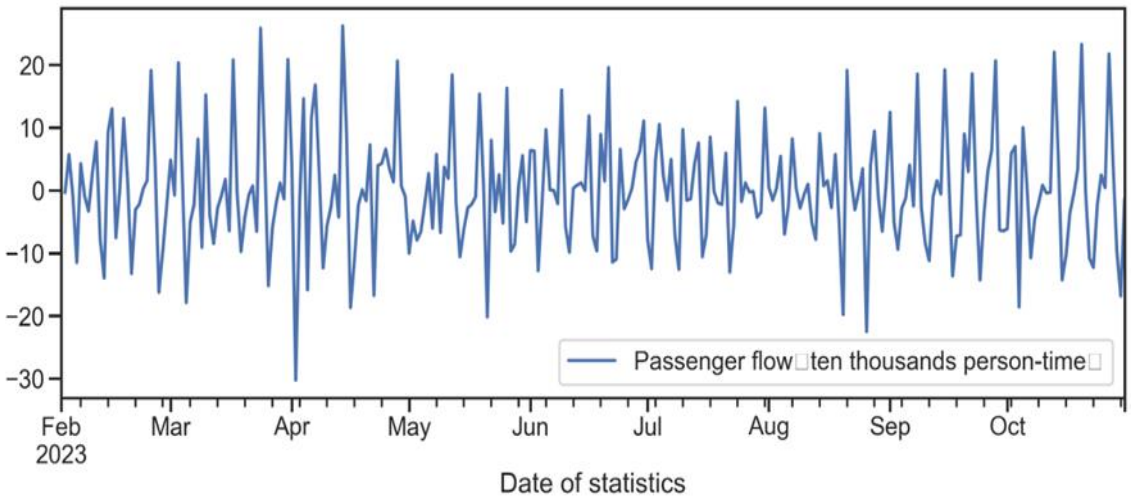

Fig. 2 Sequence diagram of training set after first-order difference

3.2. ACF and PACF Test

Differential data have stationarity, and then ACF and PACF plots are plotted to determine p and q. As shown in Figures 3 and 4.

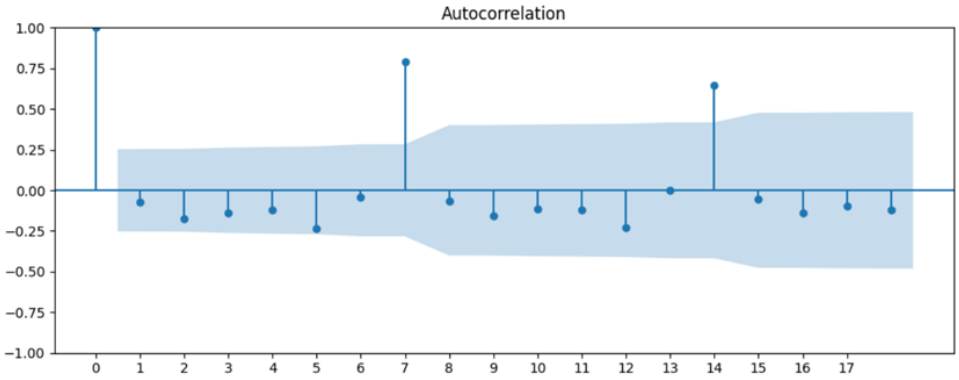

Fig. 3 ACF plot

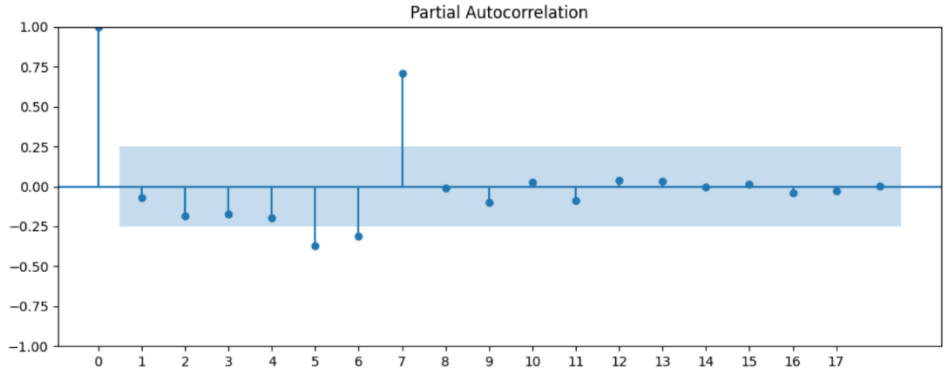

Fig. 4 PACF plot

Due to the finding that many pairs of p can be obtained by observing the ACF plot and the PACF plot, q. To determine which pair of p, the value of q is the most appropriate. The BIC thermal maps were plotted as described in Figure 5.

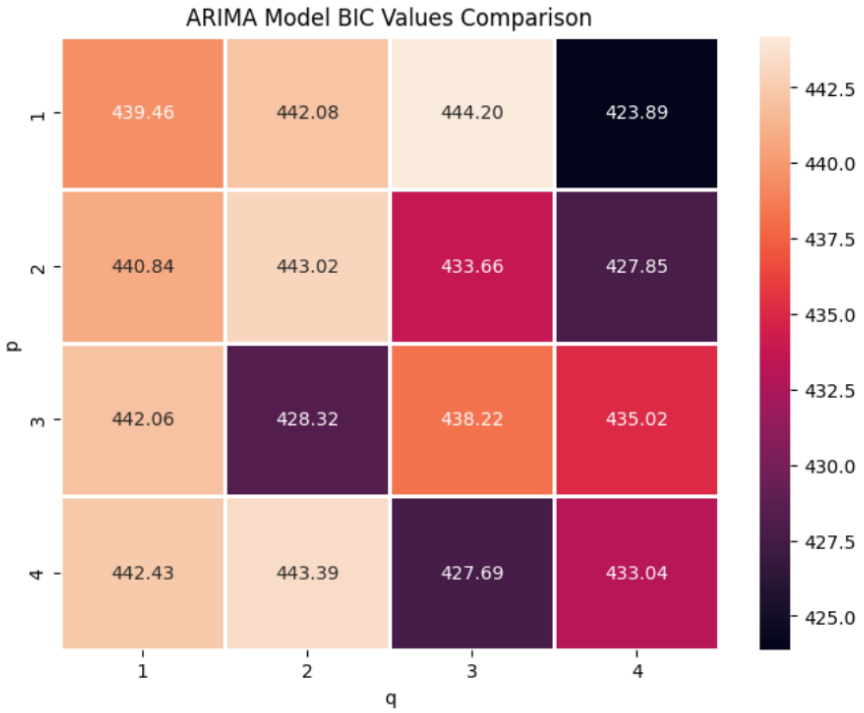

Fig. 5 BIC heatmap

The observed heat map shows that p, 1 and q, 4 are the most reasonable. The ARIMA model was constructed with (p=1, d=1, q=4) to predict the passenger flow from November 1,2023 to December 31,2023, and compare it with the true value. The available view is shown in Figure 6.

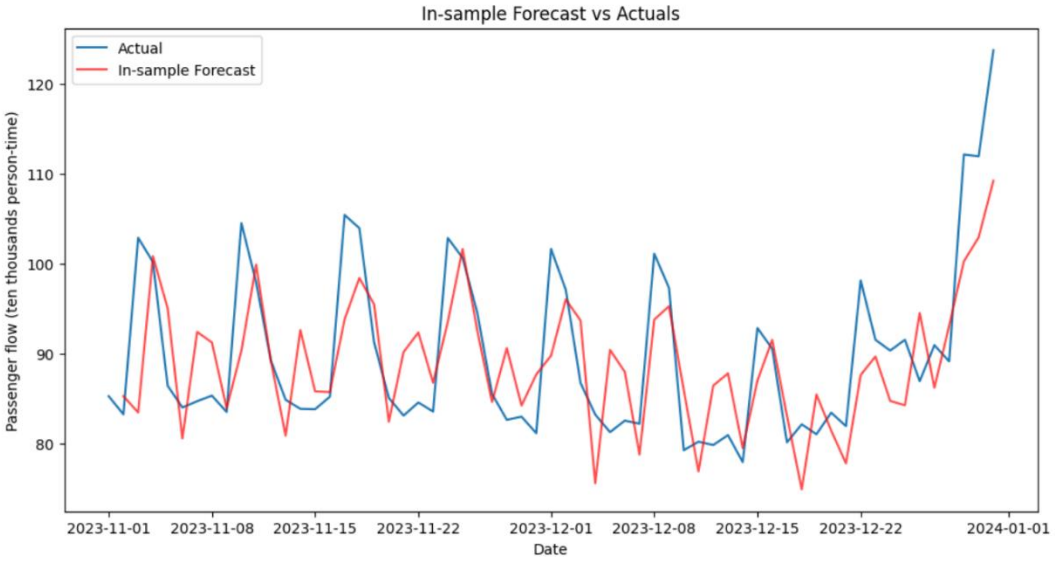

Fig. 6 The predicted value compared with the true value

Observing at the predicted value and true value. The forecast from 1 November 2023 to 30 November 2023 was more accurate, however the deviation was large from December 1,2023 to December 31,2023. Found that ARIMA is more suitable for short-term passenger flow forecast in this model and that ARIMA is more suitable for making rough prediction.

4. Conclusion

This paper used existing data to build the ARIMA model to predict the daily ridership of a single subway line. Result showed that the model has certain reliability in subway passenger flow prediction. In the research process, the author used BIC heat map to find p and q. BIC can not only find more

suitable p and q more quickly, but also save the cost of research time. Through the above demonstration, ARIMA can be well applied to the big data analysis and rough prediction of time series, however this paper also points out some limitations of this model. First, ARIMA limits the stationarity of the data, leading to the prediction of non-stationary data. Secondly, the model cannot achieve high precision prediction, can only predict the direction and rough data. Considering the large limitations of this model, future studies can combine other methods and techniques to improve the accuracy and stability of prediction or explore other time series models or combine machine learning algorithms to enhance the prediction of subway passenger flow.

References

[1] Chang Li, Zuo Zhongyi, Han Bing. The prediction of rail transit traffic based on BP neural network. Journal of Dalian Jiaotong University, 2014, 35(1): 13-16.

[2] Wen Huiying, Luo Chenwei. Short-term subway ridership prediction based on deep learning. Journal of Guangxi University (Natural Science Edition), 2020, 45(2): 389-397.

[3] Liu Chen, Chen Jingxian, Hao Yuchen, et al. Passenger flow prediction of subway in and exit stations based on space-time network. Computer Engineering and Application, 2021, 57(18): 248-254.

[4] Li Huixuan, Wu Ruiyi. Subway station passenger flow model study using the XGBoost and SVR algorithms. Journal of Sanming College, 2019, 36(6): 56-64.

[5] Tian Zhao, Cheng Yujie, Zhang Ganzhong, et al. Short-time prediction of subway passenger flow based on ASTLSTM. Journal of Zhengzhou University (Science edition), 2024, 1-7.

[6] Lu Zhiyi, Nie Weicong, Chen Lizhen. Urban rail transit traffic flow prediction based on the ARMA model. Henan Science, 2018, 36(5): 646-651.

[7] Wu Xiangbin, Liu Zhifeng, Ding Chenglong, et al. Analysis and prediction of subway passenger traffic data based on ARIMA mode. National Defense Manufacturing Technology, 2021, 4: 15-17.

[8] Shao Bilin, Rao Yuan, He Xin. Research on subway passenger traffic prediction based on SARIMA-SVM combined model. Software Guide, 2022, 21(11): 24-30.

[9] Cui Hongtao, Chen Xiaoxu, Yang Chao, et al. Forecdiction of subway passenger flow based on deep long and short-term memory network. Urban Rail Transit research, 2019, 22(9): 41-45.

[10] Peng Tongxin, Han Yong, Wang Cheng, et al. A hybrid deep learning model for short-term subway passenger traffic prediction. Computer Engineering, 2022, 48(5): 297-305.

[11] Li Mingmin, Qiu Weiyi, Jiang Shun, et al. A method for predicting the outbound passenger flow of rail transit stations based on the modified ARIMA model. Transportation and Port and Navigation, 2017, 4(2): 45-49+80.