Research on Forecast of Rail Traffic Flow Based on ARIMA Model

Shu Ying Liu, Shuo Liu, Ye Tian, Quan Long Sun, Yu Yang Tang

College of Information Engineering, Anhui Xinhua University, Hefei, AnHui, 230088, China
liushuying@axhu.edu.cn

Abstract. With the rapid economic development, subway-based rail transit is spreading all over the country, and efficient prediction of rail passenger flow is the key to alleviating traffic pressure. In view of the time-series characteristics of subway passenger flow data, the author uses the simulation results to show that the ARIMA model has higher accuracy and better effect in predicting the rail transit flow.

1. Introduction
With the rapid development of China's economy and the improvement of people's quality of life, rail transit has gradually become one of the factors that measure a city's happiness index[1]. Nowadays, urban rail transit based on subway is favored by passengers for its safety, punctuality, speed, comfort and environmental protection. Therefore, how to accurately grasp the passenger flow distribution of rail transit stations is very important for operators.

Many scholars have proposed different prediction methods for different application scenarios. There are parametric methods, non-parametric methods, and combinatorial methods. Parameter prediction mainly includes linear regression method, nonlinear regression method, filtering method, autoregressive linear process and other methods. Non-parametric prediction mainly includes non-parametric regression, neural network and other methods[2]. The combination of weighted historical average autoregressive model weighted historical average autoregressive model. ARIMA model is also a projection. Because the passenger flow of each station of urban rail transit has certain timing characteristics and periodic fluctuation trend, it meets the requirements of an ARIMA model. Therefore, this article uses days as the unit to analyze and predict the characteristics of the passenger flow of a certain station.

2. ARIMA Model
Autoregressive Integrated Moving Average Model, referred to as an ARIMA model. In the 1970s, American scholar Box and British statistician Jenkins[3] cooperated to develop the Box-Jenkins method for processing and forecasting time series data, of which the ARMA model is the main model of the method. The ARMA(p,q) model often appears in the form of ARIMA(p,d,q). It is a model established in the process of transforming non stationary time series into a stationary time series, in which the dependent variable only regresses its lag value and the present value and lag value of the random error term. The present value and lag value of the model is established by regression.

The ARIMA(p,d,q) model includes an AR(p) model and an MA(q) model, where AR(p) is autoregressive, MA(q) is moving average, p is autoregressive term; q is mobile. The number of average terms, d is the number of differences made to make the time series stationary. I am a differential model, and ARIMA is an ARMA model after differential, which ensures the stability of the data.
The basic steps of the model for prediction are as follows: first, obtain a data set, and perform stationarity detection on the data set, and perform differential processing on data that does not have stationarity. Generally speaking, after the first-order difference, the data set will be basically stable. Secondly, the values of parameters p, q are calculated by the auto-correlation function graph (ACF), partial auto-correlation function graph (PACF)[4] or heat map. Finally, substitute the calculated values of p, d, and p into the model, and train the data set to verify the model and make predictions.

3. Build an ARIMA forecast model

3.1 Data source

The data used in the experiment in this paper is the subway passenger flow of a certain city in 2015. Select a certain date from the data sample from 10:00 on January 2, 2015 to 21:00 on November 30, 2015 as the training set to predict the data from December 1, 2015 to December 7, 2015, and verify the validity of the model. Figure 1 is a graph corresponding to the training set. Through observation, it can be found that the graph has strong volatility, periodicity and regularity. It is preliminarily considered that the data set is non-stationary, and it needs to be differentially processed later. What needs to be explained here is that the most volatile position in the picture is the Spring Festival, so the passenger flow has dropped sharply.

3.2 Data stationarity detection

Stationarity refers to the fitting curve obtained through the sample time series, which can continue to follow the existing form "inertia" for a period of time in the future[5]. Stationarity requires that the mean and variance of the series do not change significantly. According to the time series scatter plot, ACF and PACF, unit root test (ADF) and other methods to determine whether the data is stationary.

Before processing the data stability, the data is reprocessed, and the data before 5 am (including 5 am) and after 22 pm (including 22 am) in the data set are deleted. The abnormal data in other time periods are replaced by the mean value method, that is, the average value of the passenger flow in the time period of the previous three days is used instead, and then the processed data is first-ordered. As
shown in Figure 2, after the first-order difference, the data has stabilized, so the value of the parameter d in the ARIMA model is 1.

![Data curve after first-order difference](image)

**Figure 2. Data curve after first-order difference**

### 3.3 p, q solution

The most critical part of the ARIMA model is to solve the parameters p and q. Generally, a heat map combined with ACF and PACF maps is used to solve the problem.

**Autocorrelation function ACF:** An ordered sequence of random variables is compared with itself. The autocorrelation function reflects the correlation between the values of the same sequence in different time series.

\[
ACF(k) = \rho_k = \frac{\text{cov}(y_t, y_{t-k})}{\text{var}(y_t)}
\]

The range of \( \rho_k \) is \([-1,1]\).

**Partial autocorrelation function PACF:** The correlation between the values of the same sequence at different time series, taking into account the values between intervals. For time series data, the partial autocorrelation function is usually unknown.

For a stationary AR(p) model, when the autocorrelation coefficient \( p(k) \) of lag k is calculated, it is actually not a pure correlation between \( x(t) \) and \( x(t-k) \). Because \( x(t) \) will also love the influence of the middle \( k-1 \) random variables \( x(t-1), x(t-2),..., x(t-k+1) \) and these \( k-1 \) Random variables are all related to \( x(t-k) \). Therefore, the autocorrelation coefficient \( p(k) \) is actually mixed with the influence of other variables on \( x(t) \) and \( x(t-k) \). Therefore, PACF eliminates the interference of the \( k-1 \) random variables \( x(t-1), x(t-2),..., x(t-k+1) \) after the interference of \( x(t-k) \) on \( x(t) \) Relevance.

The difference between ACF and PACF: ACF also includes the influence of other variables, while PACF is strictly the correlation between these two variables. Table 1 below is about how p and q are determined in ARIMA(p,d,q). Censoring: falling within the confidence interval (95% of the points meet the rule).
Table 1. The order determination table of the parameters in the ARIMA model

| Model      | ACF                               | PACF                             |
|------------|-----------------------------------|----------------------------------|
| AR(p)      | Attenuation tends to zero (tailing) | p-order after censoring          |
| MA(q)      | q-order after censoring           | Attenuation tends to zero (tailing) |
| ARMA(p, q) | Attenuation tends to zero after q order (tailing) | Attenuation tends to zero after p order (tailing) |

4. Metro data for prediction

4.1 Evaluation index
The quality of a model requires corresponding evaluation indicators for quantitative evaluation. The evaluation indicators used in this experiment are MAE and RMSE.

MAE——Mean Absolute Error. The formula is:

\[
MAE = \frac{1}{n} \sum_{i=1}^{n} |y_p - y_r|
\]  

RMSE——Root Mean Squared Error. The formula is:

\[
RMSE = \sqrt{\frac{\sum_{i=1}^{n} (y_p - y_r)^2}{n}}
\]

In the above formula, \(n\) is the number of samples, \(y_p\) is the predicted value, and \(y_r\) is the true value. The experiment obtained MAE, the smaller the RMSE value, the closer the predicted value to the true value, the better the prediction effect.

4.2 Data prediction
In this experiment, the values of \(p\) and \(q\) were obtained through the heat map. Figure 3 is the heat map obtained after the experiment. The abscissa is \(q\), the ordinate is \(p\), and the combination of small values is selected. The fitting effect of the model uses AIC and BIC as evaluation indicators.

AIC is also called Akaike Information Criterion, which is used to measure the pros and cons of analytical model fitting. Its expression is:

\[
AIC = 2k - 2\ln(L)
\]  

BIC is also called the Bayesian Information Criterion, which can effectively prevent the model from becoming too complex due to high model accuracy. Its expression is:

\[
BIC = k \ln(n) - 2\ln(L)
\]

In formula (4)(5), \(K\) is the number of model parameters, \(n\) is the number of samples, and \(L\) is the likelihood function. The evaluation criteria of BIC and AIC are the same. The smaller the value, the better the model prediction effect.
Figure 3. Heat map

It can be seen from the figure that at the point (3, 4), the value is the smallest, so the final value is $p=3$, $q=4$.

Using the model ARIMA(3,1,4) tested in the previous period to predict the time series data of subway passenger flow in a certain city, the prediction result is shown in Figure 4.

Figure 4. Comparison of predicted and actual values

In order to verify the prediction effect, this experiment constructed two models: BP model and ARIMA model. Table 2 shows the evaluation index values of the two models.

| Model | RMSE  | MAE   |
|-------|-------|-------|
| ARIMA | 1121.31 | 898.37 |
| BP    | 2607.24 | 1708.49 |
It can be seen from the above table that the RMSE and MAE values of the ARIMA model are significantly lower than the BP model, and the prediction effect is better.

5. Conclusion
This article introduces the working principle of the ARIMA model in detail, then constructs the model and performs simulation verification. After comparing with the BP model, it is found that the ARIMA model predicts better and the error is acceptable. Therefore, this model can be used to predict the flow of urban rail transit. But the ARIMA model also has certain shortcomings. ① The time sequence data is required to be stable, or the data after the difference is stable. ② The model essentially only captures linear relationships, not nonlinear relationships. Therefore, the author will further optimize the model in the later stage to improve the accuracy of prediction.

Acknowledgments
1. Quality Project of Anhui Xinhua University under Grant NO.2018 ppzyx03
2. Quality Engineering Project of Anhui NO. 2018 mooc595
3. Quality Engineering Project of Anhui Provincial Department of Education NO. 2018 jyxm1449
4. Science Research Team of Anhui Xinhua University NO. kytd201902

About the author
Liu Shuying (1978-), female, Master of Associate Professor, School of Information Engineering, Xinhua University, Anhui;
Research direction: Data analysis and mining
Mailing address: School of Information Engineering, Anhui Xinhua University, 555 Wangjiang West Road, Hefei City, Anhui Province 230088 liushuying@axhu.edu.cn; 18105692722.

References
[1] Cheng Zhaolan, Zhang Xiaoqiang, Liang Yue. Prediction of railway freight volume based on LSTM network [J]. Acta railway Sinica, 2020, 42 (11): 15-21.
[2] Li Shu, Wang Fei, Wang Fengxiao. Empirical analysis of consumer price index in Qingdao Based on ARIMA model [J]. Journal of Lanzhou University of Arts and Sciences (NATURAL SCIENCE EDITION), 2020, 34 (06): 11-16.
[3] Wang Fangfang, Wang Qian. Analysis and prediction of Shaanxi GDP Based on ARIMA model [J]. Industrial innovation research, 2020 (19): 22-24.
[4] Shuona XU, Biqing ZENG. Network Traffic Prediction Model Based on Auto-regressive Moving Average[J] Journal of Networks 2014 (23) : 201-208.
[5] Cai Hongxiu, Xu Ming. Short term forecast of international cargo throughput of Shanghai airport based on product season ARIMA model [J]. China strategic emerging industries, 2016 (28): 15-18.