Prediction of Taxi Ridership in New York City based on ARIMA Model

Xuanqi Zhang *

School of Mechanical Engineering with Transportation, Dundee International Institute of Central South University, Changsha, 410083, China

* Corresponding Author Email: 7804230130@csu.edu.cn

Abstract. With the development of society, people's travel demand is increasing. Traffic congestion has become an urgent problem to be solved. The effective prediction of passenger flow has become a key problem to solve the traffic congestion problem. There is always an accurate prediction result that allows people to better deal with the future traffic situation and make the best response strategy. In this study, the ARIMA model was used for prediction experiments. By analyzing and modeling the ridership data of New York taxi companies from January to July 2015. Get the suitable model for this study is ARIMA (3,1,1) model. And whether the data can be used by ARIMA model is tested to ensure the reliability of the model. Then, the line chart and prediction data about time and number of trips are obtained. Through this study, it can be found that ARIMA prediction model has certain practicability and reliability in predicting data. Through the analysis of the data set, the predicted value of the passenger flow from August 1 to 8 can be obtained more accurately. At the same time, there are some limitations, that can be improved when combined with other models. For example, when dealing with the data set related to vehicles in this study, the ARIMA model could not be used for predictive analysis because the data set itself had many extreme values and outliers.

Keywords: Taxi ridership; ARIMA model; white noise hypothesis.

1. Introduction

With the continuous development of industrialization and the progress of science and technology. People's pace of life is getting faster and faster. More and more people are not satisfied with the traditional way of travel. The emergence of cars, subways, trains and planes has gradually enriched people's travel. People have more choices and a variety of travel methods will facilitate communication and economic development between people in different regions. At the same time of rapid development, there are also some problems. On the ground, due to the increase of vehicles and the increase of people's travel demand, traffic congestion is a problem. At the same time, because of the concentration of people's travel time, whether by subway or by car will cause congestion. Therefore, the correct prediction of passenger flow is a very important problem to be solved.

Passenger flow forecasting has always been a very important experiment and research topic in the development of urban rail transit. Only by accurately predicting the passenger flow in the future period can people prevent the traffic congestion problem more effectively. There are two common passenger flow forecasts. One is short-term passenger flow forecast, and the other is medium-term passenger flow forecast. Short-term forecasting is usually based on a period of time from 15 minutes to 60 minutes for the statistics and analysis of data. It can reflect the change of passenger flow in a day. Most of the medium-term passenger flow forecast is based on the statistics and analysis of data per day [1]. This can more clearly reflect the change of passenger flow in a longer period of time.

Nowadays, people usually use the method of machine learning to predict passenger flow. However, artificial neural network, support vector machine and K-nearest neighbor algorithms commonly used in machine learning have some limitations in passenger flow prediction [2]. Compare with the above methods, Autoregressive Moving Average Model (ARIMA) model and The Long Short-Term Memory (LSTM) model can analyze data more effectively and obtain prediction results more accurately [3]. They are commonly used in research.

The Long Short-Term Memory (LSTM) is generated to solve the gradient disappearance and gradient explosion of Recurrent Neural Network (RNN). When processing data, you can selectively

forget some of the less important information, so that you can remember a longer sequence [4]. Li combined the improved Particle Algorithm (IPSO) with the LSTM model, and proposed the IPSO-LSTM rail passenger volume prediction model [5]. Ding et al. determined that LSTM has excellent performance in considering factors such as weather, season, precipitation, holidays, and user behavior through the learning of long and short neural memory network prediction, and has good performance in forecasting [6].

Autoregressive Moving Average Model (ARIMA) is also a commonly used time series analysis method. It uses the concepts of autoregression and moving average to model the trend and randomness of time series data [7-9]. Bao et al. predicted short-cycle traffic flow at traffic intersections by combining ARIMA and Convolutional Neural Network (CNN). It is found that combining ARIMA improves the accuracy of prediction results. And the algorithm can maintain high accuracy regardless of data set size [10]. The following research will use the ARIMA prediction model as a tool. Analyzing and forecasting the ride data of a car rental company in New York.

2. Methods

2.1. Data Source

This study uses the accurate and objective data source from New York, USA. The data is measured in days and contains 26,182 entries from January to July. Table 1 shows the full names, data types, and explanations of the three variables used in the study. Counting the date can make people more clearly see the change of passenger flow over time, and the data is more accurate and authoritative. The number of trips is a good indication of the number of passengers in a day. The number of vehicles shows how many cars the company sent out that day to cope with the day's ridership.

Table 1. Explanation of variables

Full Name	Data type	Explanations
Pick up date	Typedef	The date of receiving the guests
Number of trips	INT	Number of orders
Number of vehicles	INT	Number of cars

2.2. Method Introduction

The Autoregressive Integrated Moving Average Model is also called ARIMA model. When analyzing data, it is usually presented in the form of ARIMA (p, d, q). It is common to see the trace of ARIMA model when analyzing time series.

The three parameters: p, d, q. have important function in the model. P stands for the number of autoregressive terms. D is the number of differences. Q is used to indicate the number of moving average terms. The ARIMA is mathematically expressed as:

$$\widehat{y_t} = \mu + \varphi_1 * y_{t-1} + \cdots + \varphi_p * y_{t-p} + \theta_1 * e_{t-1} + \cdots + \theta_q * e_{t-q} \quad (1)$$

In this, the φ table shows the series of AR, and the θ table shows the series of MA where φ represents the coefficient of AR and θ represents the coefficient of MA.

In order to facilitate the calculation and statistics of the data, the data set is organized. The author plus the number of trips and the number of vehicles from different locations on the same day together in order to construction ARIMA models and prediction on a daily basis.

3. Results and Discussion

There are three parameters in the data. Therefore, this study should establish two ARIMA models to predict and analyze number of trips and number of vehicles respectively. Find changes and trends in the data, and think about the predicted value. The study will analyze the data about vehicles first.

3.1. Number of Vehicles Model Fitting

Table 2 shows that the Constant item has a sign, the Autoregressive model has two signs and the Moving average model has three signs. The number of signs corresponds to the values of d,p, and q in the ARIMA model. So the optimal model was ARIMA (2,1,3). Through the analysis of Coefficient, Standard Error, Z value, P value and 95%CI in the table, its formula was obtained as follows:

$$y_t = -25.462 + 1.239y_{t-1} - 0.995y_{t-2} - 1.727\varepsilon_{t-1} + 1.556\varepsilon_{t-2} - 0.488\varepsilon_{t-3} \quad (2)$$

Table 2. ARIMA (2,1,3) model parameter for vehicles prediction

Item	Sign	Coefficient	Standard Error	Z value	P value	95%CI
Constant	c	-25.462	23.261	-1.095	0.274	-71.053 ~ 20.128
AR	α1	1.239	0.010	126.404	0.000	1.220 ~ 1.258
	α2	-0.995	0.009	-112.483	0.000	-1.012 ~ -0.978
MA	β1	-1.727	0.068	-25.582	0.000	-1.860 ~ -1.595
	β2	1.556	0.104	14.946	0.000	1.352 ~ 1.760
	β3	-0.488	0.067	-7.241	0.000	-0.620 ~ -0.356

According to the results of Q statistics (Table 3), if the P-value of Q6 is less than 0.05, the null hypothesis is rejected at the significance level of 0.05, and the residual of the model is not white noise, and the model residual does not meet the requirements.

Table 3. Model Q statistics table for vehicles prediction

Item	Statistic	P value
Q1	1.520	0.218
Q2	15.997	0.000**
Q3	21.676	0.000**
Q4	24.440	0.000**
Q5	39.091	0.000**
Q6	39.846	0.000**
Q7	74.398	0.000**
Q8	74.876	0.000**

3.2. Number of Trips Model Fitting

Based on the Number of Trips and AIC information criterion (the lower the value is, the better), multiple potential alternative models were modeled and compared. Finally, Table 4 shows the Constant item has a sign, the Autoregressive model has three signs and the Moving average model has one sign. And through the analysis about four-term parameters the same as Table 2. The ARIMA model (3,1,1), and its formula was as follows:

$$y_t = -67.139 + 0.439 * y_{t-1} - 0.178 * y_{t-2} - 0.162 * y_{t-3} - 0.608 * \varepsilon_{t-1} \quad (3)$$

Table 4. ARIMA (3,1,1) model parameter for trips prediction

Item	Sign	Coefficient	Standard Error	Z value	P value	95%CI
Constant	c	0.959	4.541	0.211	0.833	-7.941 ~ 9.859
AR	α_1	0.439	0.168	2.604	0.009	0.108 ~ 0.769
	$\alpha 2$	-0.178	0.076	-2.348	0.019	-0.327 ~ -0.029
	α_3	-0.162	0.096	-1.690	0.091	-0.349 ~ 0.026
MA	β_1	-0.608	0.157	-3.867	0.000	-0.917 ~ -0.300

The absence of autocorrelation in residuals is a fundamental requirement in the ARIMA model. This is also known as white noise. And it can be text through the detection of q. In common cases, q_6 can always show whether the requirements are met. If the assumptions are met, it usually means that the model is working properly. It is clearly shows in Table 5, the model ARIMA (3,1,1) basically meets the requirements.

Table 5. Model Q statistics table for trips prediction

Item	Statistic	P value
Q1	0.005	0.945
Q2	0.372	0.830
Q3	1.802	0.614
Q4	2.183	0.702
Q5	2.693	0.747
Q6	2.922	0.819
Q7	17.758	0.013*
Q8	20.309	0.009**

3.3. Prediction Results

Figure 1 is a line chart about pigs, from which it can be clearly found that there is a relatively obvious deviation between the actual value and the predicted value, so the predicted value obtained from this is not much reference value.

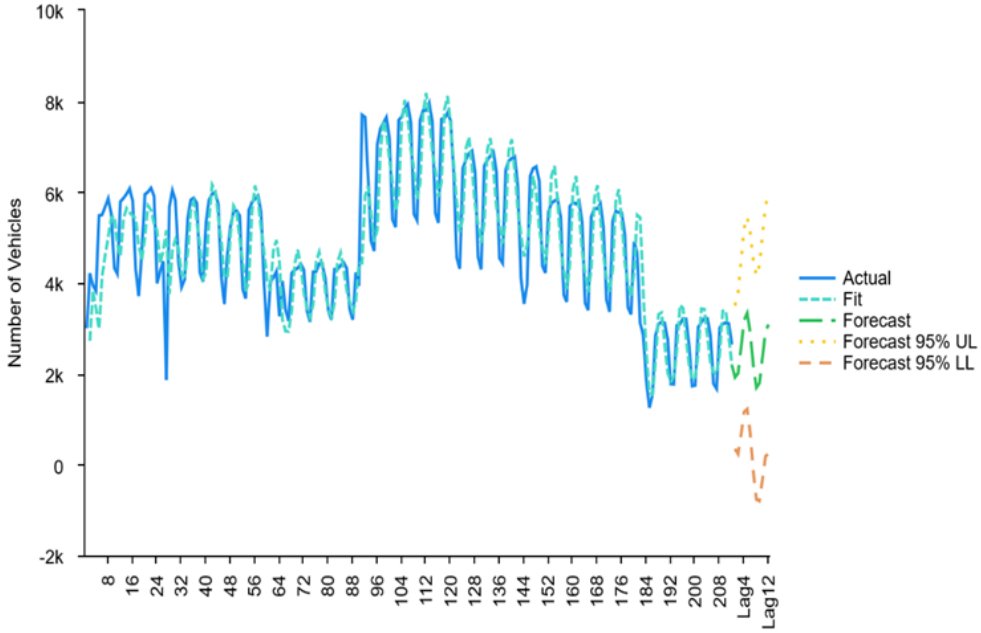

Fig. 1 Number of Vehicles' model fitting and analysis

As can be seen from Figure 2, the predicted value has a high degree of fitting and is very close to the true value, which can be used for later data prediction and observation of changes.

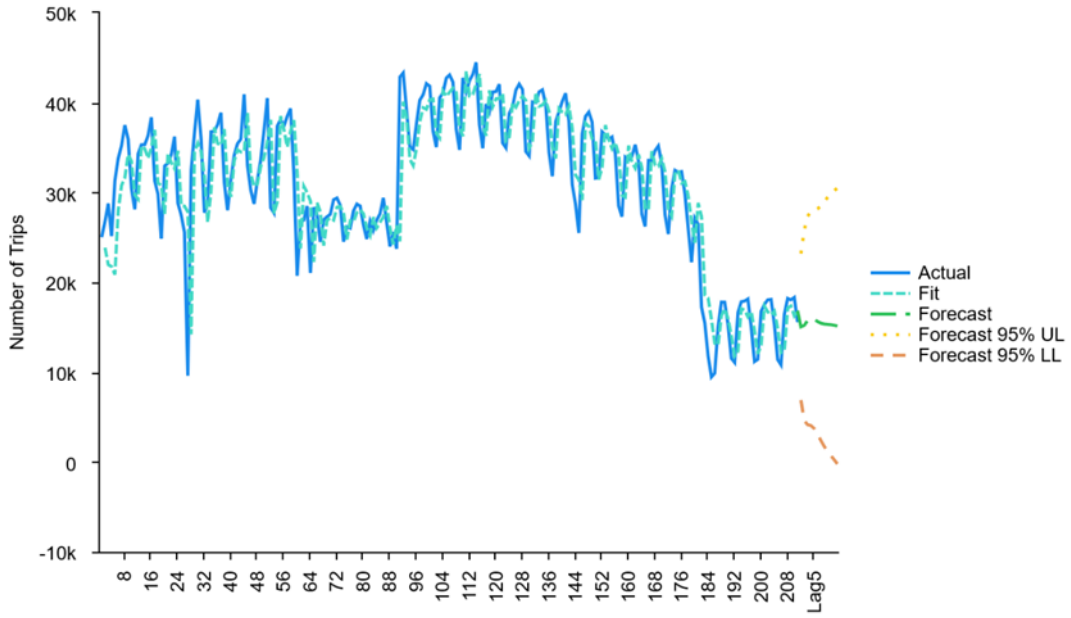

Fig. 2 Number of Trips' model fitting and analysis

Table 6 shows that the number of trips in the first four days of August has an upward trend and reaches the maximum on the fourth day. After the 4th, the number of trips decreased slightly. The overall data did not show significant changes, but it can be seen that ARIMA model has a good effect on data prediction.

Table 6. Ridership prediction

Prediction	T=1	T=2	T=3	T=4	T=5	T=6	T=7
Value	15050.892	15213.081	15752.327	15973.832	15881.360	15646.923	15457.501

4. Conclusion

The purpose of this study is to discuss the validity and reliability of ARIMA model in predicting taxi ridership. Through the analysis and modeling of the passenger volume of a taxi company in the United States from January to July, the potential and application of ARIMA model in data forecasting is demonstrated.

First, the research data were sorted and pre-processed to ensure that the data were arranged in an orderly manner according to the time series. To ensure that the data can be efficiently run in the ARIMA model. Then, according to the characteristics of time series data, the appropriate model is selected to determine the values of p, d and q parameters. Make predictions about future data after ensuring that the data meets the white noise hypothesis. At last, analyzing the results.

In the forecast and analysis of the future taxi passenger volume, it is known that the change in the next few days is relatively smooth, without too much twists and turns. And found that the number of people taking taxis since July has decreased significantly compared with April and May. The company can arrange the right number of taxis to work according to the forecast results, while ensuring benefits while minimizing costs.

However, the author also found some limitations of the ARIMA model. The ARIMA model is unable to perform effective analysis of vehicle-related data. This also shows that the ARIMA model does not perform well when dealing with data sets with many extreme and outlier values.

All in all, ARIMA model has relatively good practicability in predicting data and has certain practical potential in analyzing practical problems. But there are also some limitations. In future

research, this paper can consider combining ARIMA model with other models to make up for the problems of ARIMA model.

References

[1] Liu Y, Yu H, Fang H. Application of KNN prediction model in urban traffic flow prediction. In: Proceedings of the 2021 5th Asian Conference on Artificial Intelligence Technology, 2021, 389-392.

[2] Xu D, Wang Y, Peng P, et al. Real-time road traffic state prediction based on kernel-KNN. Transportmetrica, 2020, 16(1): 104-118.

[3] Rahman F I. Short-term traffic flow prediction using machine learning KNN, SVM, and ANN with weather information. International Journal for Traffic & Transport Engineering, 2020, 10(3): 8.

[4] Ji X, Ge Y. Holiday highway traffic flow prediction method based on deep learning. Journal of System Simulation, 2020, 32(06): 1164-1171.

[5] Li W, Feng F, Jiang Q. Improved particle swarm algorithm to optimize the LSTM neural network rail passenger traffic prediction. Journal of Railway Science and Engineering, 2018, 3274-3280.

[6] Ding Z, Tang G, Zhang B, et al. Traffic flow prediction based on improved long short-term memory neural network. Network Security and Data Governance, 2023, 42(08): 52-58.

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

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

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

[10] Bao W, Liu H, Fang Y, et al. Short-term traffic flow prediction at intersections using CNN-ARIMA. Automotive Practical Technology, 2023, 48(22): 178-182.