Prediction of High-speed Traffic Flow in Baoan District in Shenzhen

Yanqing Lu *

School of Information Science and Technology, Shanda University, Shanghai, 201209, China

* Corresponding author: 1715010123@stu.hrbust.edu.cn

Abstract. Nowadays, with the increasing number of cars, high-speed congestion has become a common phenomenon. This paper studies the traffic flow prediction of an expressway in Shenzhen, aiming to avoid congestion at the expressway during holidays and carry out effective management. The author uses the ARIMA model for research, with data selection of vehicle flow at expressway intersections from December 2002 to June 2023. Firstly, the ADF test is used to check the feasibility of the data. The ACF and PACF are used to determine the values of p and q in the ARIMA model. What is more, the optimal model is selected by comparing the BIC values of the models constructed with different p and q values. Finally, model establishment and analysis are carried out. It is found that in the 14 years after the selected data, the traffic flow is still on the rise, and it is necessary to pay attention to the traffic flow management of this high-speed section.

Keywords: ARIMA model; highway; traffic flow.

1. Introduction

Since the 21st century, the number of people owning cars has been increasing, but at the same time, it has also brought about the problem of rapid increase in traffic flow on highways [1]. Nowadays, especially on holidays, traffic congestion will always occur on the highway, so it is important to forecast the future traffic flow because it can help to manage the traffic on the highway.

Making a good traffic flow prediction can help people better understand road conditions and how to choose the best travel time. There are generally three methods of prediction: qualitative prediction, time series prediction, and causal model prediction. The prediction of traffic flow on highways is a non-linear stochastic process influenced by many fuzzy and uncertain factors, such as climate, season, economy, etc. Using qualitative and causal prediction methods for traffic flow prediction is not ideal [2]. The article uses the Autoregressive Integrated Moving Average (ARIMA) Model. The passenger flow of urban rail transit exhibits distinct temporal characteristics, namely the dynamic pattern of passenger flow over time. Due to its susceptibility to various factors, the magnitude of passenger flow changes in each cycle varies and demonstrates non-stationary time series characteristics. Consequently, previous studies have proposed ARIMA-RBF combined prediction models by integrating neural networks with the ARIMA Model [3]. A systematic analysis is conducted on ARIMA model recognition, model validation, and model prediction, which is applied to monthly traffic volume prediction on a certain expressway. The application results show that the comprehensive error rate of the model prediction is lower than the error rate of the grey model [4]. The ARIMA prediction model can better adapt to monthly traffic volume prediction on highways [4]. Some scholars also use other methods to predict the traffic flow on highways, like the traffic flow prediction at intersections based on the Gaussian radial basis function neural network [5], the highway traffic noise prediction model based on equivalent vehicle flow [6], and the highway traffic flow prediction based on grey prediction model [7].

The ARIMA model in this paper collects daily traffic flow and analyzes it to obtain predictions for future traffic flow. The traffic flow data of more than 200 days of high-speed intersections are selected for analysis, to reduce the occurrence of traffic congestion and make management resources more reasonable allocation. The ARIMA model comprehensively considers the changing trend of station passenger flow, including boarding and alighting passenger flow, as well as other interference factors. It has the advantages of easy quantification and high accuracy [8].

Some other articles aim to construct an ARIMA model for predicting traffic flow, which involves assessing the stationarity of the time series, determining the appropriate model order, and conducting a comprehensive analysis [9]. Part of them contend that this kind of study bears significant implications for the development of urban transportation, as it can alleviate strain on high-speed transportation systems, facilitate the implementation of more rational traffic management strategies, and contribute to fostering sustainable city development [10].

The article aims to solve the current situation of frequent traffic jams on highways, through the traffic flow forecasting model, it can reduce the occurrence of this kind of situation, and the traffic is guaranteed safe and smooth.

2. Methods

2.1. Data Source

The dataset selected by the author comprises traffic volume data on highways in Baoan District from December 2022 to June 2023, obtained from the Shenzhen Open Data Platform. These data are categorized daily and possess high authenticity for predicting vehicle flow patterns. Subsequently, the author performed partial deletion of irrelevant columns, retaining only the time and passenger flow variables.

2.2. Model Introduction

The ARIMA model aims to uncover latent time series patterns by leveraging autocorrelation and differentiation and subsequently utilizes these patterns for future data prediction. The autoregressive (AR) component is employed to handle the influence of past observations on the current value within the time series. Meanwhile, the moving average (MA) component deals with the impact of previous prediction errors on the present value. By integrating these three components, the ARIMA model effectively captures trend changes in data and effectively handles temporary fluctuations, abrupt shifts, or high levels of noise. The mathematical formulation of the ARIMA model can be expressed as follows:

$$Y_t = c + \varphi_1 Y_{t-1} + \cdots + \varphi_p Y_{t-p} + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} + \varepsilon_t \quad (1)$$

In the ARIMA (p, d, q) Model, q is for Moving Average, p is for Autoregressive and d is for Difference Order. The difference order needs to be calculated to make the target sequence stationary, here is the formula for the calculation:

$$\mathrm{y_t} = \mathrm{y_t} - \mathrm{y_{t-n}} = \mathrm{y_t} - \mathrm{B^n y_t} = (1 - \mathrm{B^n})\mathrm{y_t} \quad (2)$$

The Augmented Dickey-Fuller test (ADF) is a widely employed statistical test for assessing the stationarity of a given time series, making it one of the most frequently utilized statistical tests in sequence stationarity analysis, the ADF test is fundamentally a statistical significance test, this means that there is a hypothesis test involving the null hypothesis and the alternative hypothesis, so the test statistic is calculated and a p-value is reported. Based on the test statistics and the p-value, you can infer whether a given sequence is stationary, and here is the formula for the calculation:

$$y_t = c + \beta t + \alpha y_{t-1} + \varphi_1 \Delta Y_{t-1} + \cdots + \varphi_p \Delta Y_{t-p} + e_t \quad (3)$$

The Autocorrelation Function (ACF) employed in this study serves as a quantitative measure to assess the correlation between a time series and its lag, while the Partial Autocorrelation Function (PACF) is utilized to evaluate the correlation specifically between a time series and its lag. Through ACF and PACF graphs, the autocorrelation and partial autocorrelation of time series can be analyzed, to provide a basis for the establishment of time series models.

3. Results and Discussion

3.1. Stationary Test

Figure 1 shows the original data, but it isn't easy to judge whether the data is stable, so it is necessary to conduct an Augmented Dickey-Fuller test (ADF).

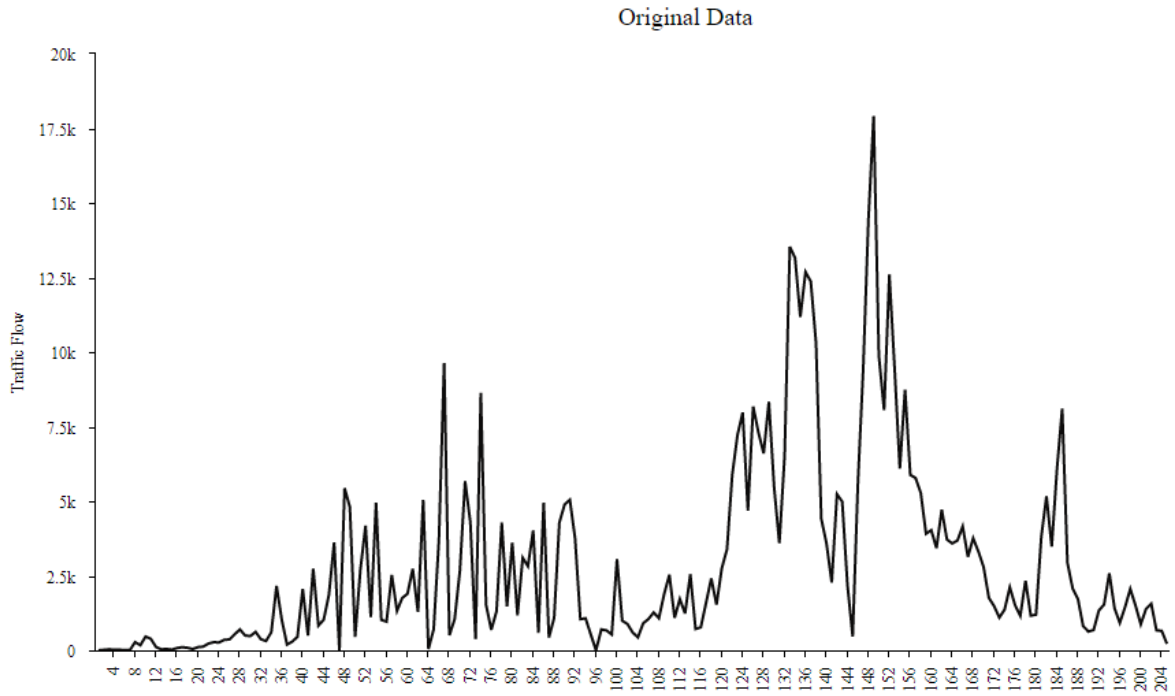

Fig. 1 Time series plot

According to Table 1, after ADF detection, it is found that the P-value is 0.013, and the standard for judging sequence stationarity is that the P-value is less than 0.05. The p-value satisfies this condition, so the d value in the ARIMA (p, d, q) model is 0.

Table 1. ADF test table

Difference order	t	p	critical value		
			1%	5%	10%
0	-3.353	0.013	-3.463	-2.876	-2.574

3.2. Model Building

The Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) are employed in this article to determine the optimal values of p and d for the ARIMA (p, d, q) model. In the ACF diagram, the X-axis represents the lag and the Y-axis represents the autocorrelation coefficient. For each lag number, the graph has a vertical line representing the autocorrelation coefficient, the value of which is the correlation between the time series and its lag version. In the PACF diagram, the X-axis represents the lag number and the Y-axis represents the partial autocorrelation coefficient. Through Figure 2 and Figure 3, it is found that the data tends to be stable when no difference is made, so the values of p and q can be selected as 0, 1, 2.

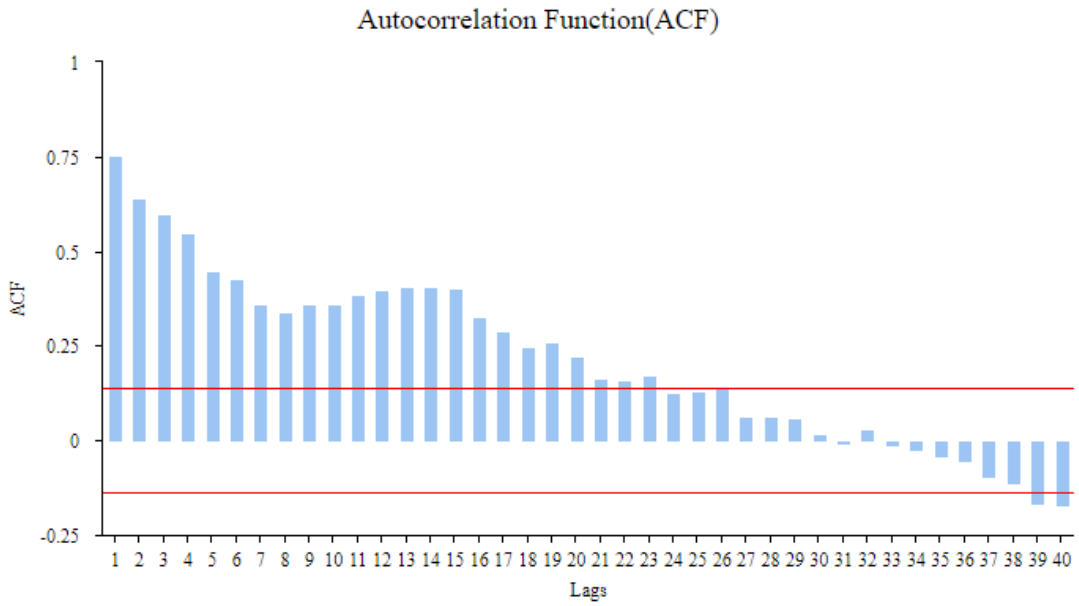

Fig. 2 Results of ACF

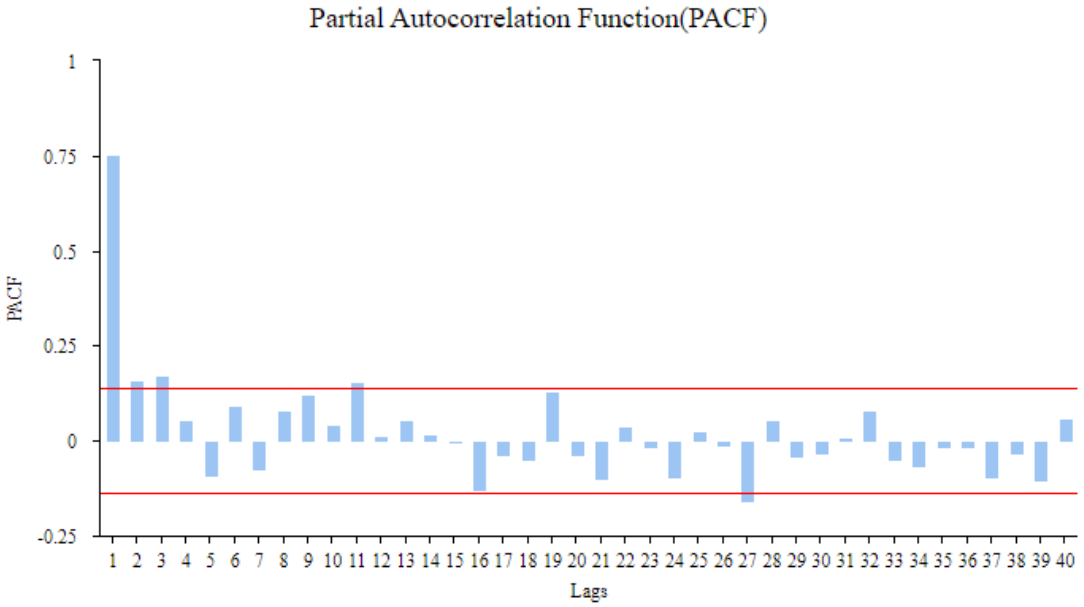

Fig. 3 Results of PACF

Root Mean Square Error (RMSE) can represent the average deviation of the disparity between the anticipated value and the actual value, in this article, it is used to select some models with less error. What is more, each model has its corresponding Bayesian Information Criterion (BIC) value, and the smaller the value, the better the choice of model, the following is the calculation formula for BIC:

$$BIC = -2\ln(L) + \ln(n) * k \tag{4}$$

Table 2 records the BIC and the RMSE values of the model composed of each p and q value that the article selects, and some models have similar RMSE values like the ARMA (1, 1) model and the ARMA (2, 2) model, their margin of error is small, so the BIC values can help to choose the better model. It is found that the ARMA model (1, 1) has the lowest BIC values, so this model is the best choice.

Table 2. Results of BIC and RMSE values

	RMSE	BIC
ARMA(0,0)	3219.588	3903.984
ARMA(1,0)	2108.605	3735.581
ARMA(1,1)	2060.876	3731.651
ARMA(0,1)	2522.419	3809.290
ARMA(0,2)	2323.982	3780.970
ARMA(2,0)	2079.803	3735.270
ARMA(1,2)	2063.344	3737.057
ARMA(2,1)	2068.823	3738.152
ARMA(2,2)	2053.806	3740.720

From Figure 4, the dashed line represents the predicted value, and the implementation represents the raw data. It can be seen that the trend of the fitted graph is consistent with the actual value, so the ARMA (1, 1) model is feasible. Compared with models with other parameters, this model has the highest degree of fit.

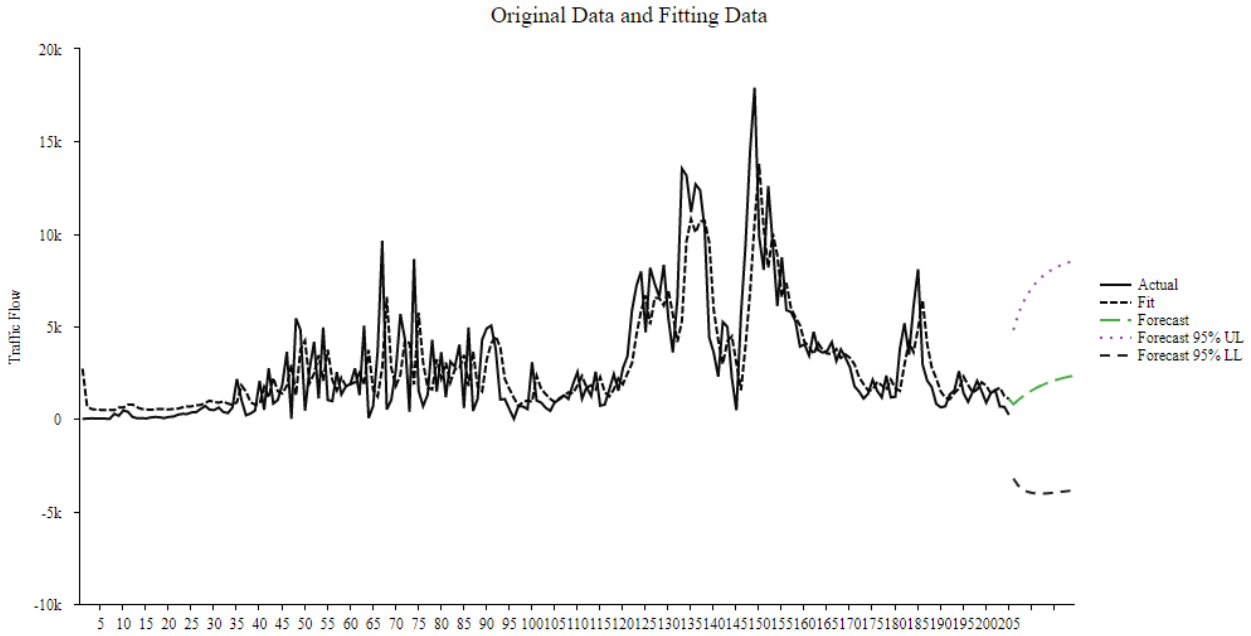

Fig. 4 Model fitting results

3.3. Model Prediction

This paper predicts the results of the selected data for the next 14 periods. From the data in Table 3, the traffic flow is trending upward. Therefore, through the prediction, it means that there is a high possibility of traffic jams after the highway section, so it is necessary to strengthen traffic flow control at this high-speed intersection in the future.

Table 3. Forecast results

Forecast	Value	Forecast	Value
Leg1	788.603	Leg8	1899.174
Leg2	1010.917	Leg9	1993.564
Leg3	1207.623	Leg10	2077.081
Leg4	1381.670	Leg11	2150.978
Leg5	1535.670	Leg12	2216.363
Leg6	1671.931	Leg13	2274.217
Leg7	1792.496	Leg14	2325.406

4. Conclusion

The article uses the ARMA (1, 1) model, compared with the values of other parameters, it is found that only the (1, 1) model has the best fitting degree, and the values of Akaike Information Criterion (ALC) and Bayesian Information Criterion (BLC) are the smallest. Therefore, for the research on this data, the choice of the model in the paper is optimal. What is more, this article also identifies some limitations in the analysis of the dataset under consideration. The daily data is employed in this study, while the hourly vehicle flow dataset is incorporated to enhance the predictive precision. Furthermore, the integration of other models with the ARIMA model used in this study will make more favorable outcomes.

References

[1] Sun Xiaojun, Cao Luling, Pan Hongmei, et al. A new model for predicting freeway traffic flow. Computing technology and automation, 2016, 35(2): 4.

[2] Zhang Chen, Du Junping. Research on the forecasting method of freeway traffic flow. Journal of Beijing Technology and Business University: Natural Science Edition, 2001, 19(4): 4.

[3] He Jiuran, Si Bingfeng. Application of ARIMA-RBF model in urban rail transit passenger flow prediction. Shandong science, 2013, 26(3): 7.

[4] Rui Shaoquan, Kuang Anle. Monthly highway traffic ARIMA forecast model. Journal of Changan University: Natural Science Edition, 2010, 30(4): 5.

[5] Tang Yan, Wang Hongbo, Wang Wanxin. Traffic flow prediction at intersection based on Gaussian radial basis function neural network. Agricultural equipment and vehicle engineering, 2006, 3: 3.

[6] Li Zezheng, Xu Jiangrong, Zhai Guoqing. Highway traffic noise prediction Model based on equivalent vehicle flow. Engineering Construction and Design, 2008, 9: 4.

[7] Wang Chao, Sun Weihua, He Yuanlie. Application of grey prediction model in expressway traffic flow prediction. Journal of Guangdong University of Technology, 2012, 29(1): 3.

[8] Zhang Bomin. Study on short-term forecast of passenger flow of Shanghai-Nanjing intercity railway. Chinese railway, 2014, 9: 6.

[9] Sun Min. Traffic flow analysis and prediction of Chongzun expressway based on ARIMA model. Shandong Communications Technology, 2014, 3: 2.

[10] Zou Aijuan, You Zilin, Wu Dan. Short-term forecast of expressway traffic volume based on ARIMA model. Defense transportation engineering and technology, 2015, 13(6): 4.