Traffic Prediction of Urban Intersections based on ARIMA Model

Huihui Chen *

Department of Information Management, Shanghai Lixin University of Accounting and Finance, Shanghai, 201209, China

* Corresponding Author Email: 221650310@stu.lixin.edu.cn

Abstract. In the urban planning, road traffic planning as an important part, through the formation of traffic layout network, can further accelerate the pace of urban development. However, the blockage will cause the car to start and stop regularly, and the car engine will emit a large amount of exhaust gas in these two stages, resulting in increased environmental pollution. Therefore, the focus needs to be on how to predict traffic flow efficiently. This paper focuses on the traffic flow density of intersections in one month, presents the change of traffic density by means of mapping, reasonably divides the training set and the test set, and uses ARIMA model to make effective prediction. It shows the changing relationship between traffic flow and time. Through reasonable prediction, the congestion problem is found in the first time, and at the same time, the solution measures are given, the traffic infrastructure is scientifically allocated, and the safety control system is set up, which provides reference for the later urban traffic development.

Keywords: Traffic flow; forecasting; ARIMA model.

1. Introduction

With the continuous acceleration of the urbanization process and the gradual improvement of people's daily living standards, the number of motor vehicles has maintained a steady upward trend, which has led to the increasingly severe urban traffic congestion. In order to alleviate traffic congestion, traffic managers mainly start from the aspect of traffic demand, adopt advanced traffic management means, scientifically, reasonably and effectively use vehicles, make full use of existing road resources, and improve the service level of roads [1]. The traffic flow can be predicted one day or more in advance, it can not only help managers arrange deployment, reasonable traffic diversion and evacuation in advance, but also select travel routes for the public to refer to in advance, which is also an effective path to alleviate highway congestion [2]. Traffic flow prediction is an important part of intelligent transportation system, which is of great significance for traffic management, route planning, reducing traffic congestion and improving road efficiency.

Traffic flow prediction is the process of analyzing the road traffic conditions of urban traffic [3]. It can not only predict the future traffic flow, vehicle speed and vehicle density on the road, but also realize the excavation of road traffic patterns and the prediction of the future traffic trend of the urban road network [3]. Intelligent transportation is the main development trend of modern transportation, which uses a large number of intelligent technologies and intelligent devices to realize the transmission and sharing of data and the utilization of value [4]. For example, some researchers use Linear Regression to establish the relationship between traffic flow and other variables (such as weather, traffic accidents, etc.), while others use Auto-Regressive Moving Average (ARMA) to process time series data.

Liu et al. found that the Autogressive Integrated Moving Average (ARIMA) model can effectively identify holiday passenger flow characteristics, reduce the workload of passenger flow prediction in the early stage, and realize the passenger flow prediction at various periods of urban rail transit holidays [5]. Zhang et al. found that the ARIMA model had better forecasting effect in the short-time passenger flow prediction of urban rail transit system [6]. Therefore, this paper mainly uses ARIMA to predict the traffic flow at intersections, which is helpful for traffic management and planning, reducing traffic congestion and improving road use efficiency. ARIMA model is a common time

series forecasting model with simple structure [7]. ARIMA model can deal with non-stationary time series, traffic flow is non-stationary in time series, its mean and variance time change with time, and ARIMA can convert non-stationary time series into stationary time series [8]. ARIMA models can be used to process non-stationary time series by automatic difference, making them stationary to satisfy the assumptions, and at the same time relying only on the historical data of the time series itself, without the need for additional external explanatory variables. In most cases, ARIMA is able to provide relatively accurate short-term forecasts.

However, when the time series data has fluctuations, trends and irregular components, the prediction accuracy will be affected without decomposition [9]. The data are cleaned first (such as removing or filling missing values, processing outliers, etc.) in this paper, and then the stationary detection and white noise detection are carried out. When the residual of the data meets the normal distribution and there is no autocorrelation, the ARIMA model is used to predict the time series data. Through the combination of real-time detection data and historical data, the traffic situation is analyzed and predicted, and the results are timely fed back to the road management department, so as to achieve reasonable guidance and regulation of road vehicles [10].

2. Methods

2.1. Data Source

The data in this paper are from a two-lane intersection in a city in Afghanistan, and the data information is collected by computer vision models. The data throughout the whole day is taken as the sample, which is from December 10, 2023 to January 9, 2024. The data mainly includes the Date, Time, Day of week, Car Count, Bike Count, Bus Count, Truck Count, Total and Traffic Situation. This paper mainly researches the effect of time on traffic flow, so only four data are used in this paper: Date, Time, Total, and Traffic Situation, are shown in Table 1.

Table 1. Information of initial data

Column Name	Explain	Example
Date	Take one day of the month as the sample	2023/12/10
Time	Time slices at 15-minute intervals	0:00:00
Total	The total number of vehicles monitored by the model	39
Traffic Situation	The condition of urban intersections density	Low

In order to improve the quality and availability of the data and improve the performance of the model and the accuracy of the prediction, it is necessary to forecast the data first. Data quality is improved and data is converted into a reasonable format by cleaning missing values, outliers, duplicates, and inconsistencies. The simplified data set can reduce the overfitting of the model to the training data, and make the model more generalized and better able to deal with the unknown data. The average of the total number of vehicles on that day is used to fill the null value in the data set in this paper, so as to obtain a complete data set.

After data preprocessing is completed, the frequency distribution diagram of traffic conditions is made, is shown in Figure 1. The histogram shows the frequency of different traffic conditions. From Figure1, it can be seen that the traffic conditions are normal in most of the time periods, but the traffic conditions are still high and heavy in some of the time periods. Thus, the traffic flow under the condition of high congestion and heavy congestion should be paid attention to.

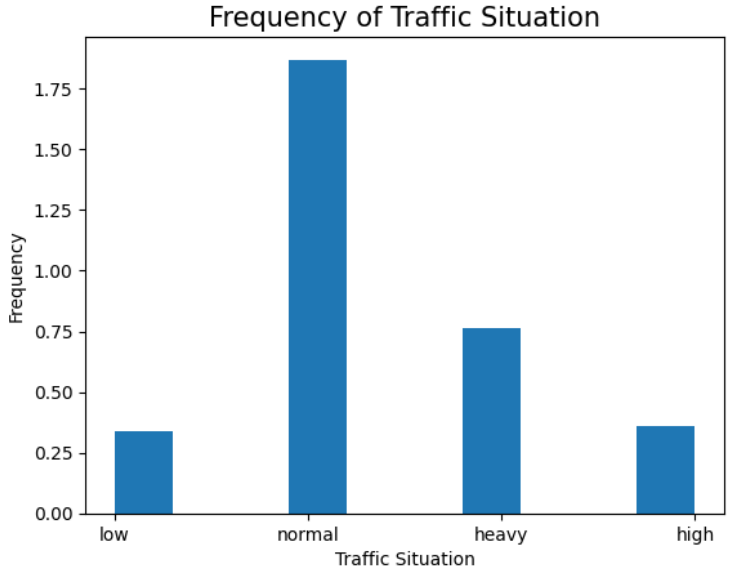


Fig 1. Frequency of Traffic Situation

The time of the data set is continuous and uninterrupted. For better prediction in the future, this paper sliced the time, taking 15 minutes as slice, and calculates the data every 15 minutes, is shown in Table 2. The one-month time series data will be used as a training set from December 15 to January 4, and as a test set from January 5 to January 9.

Table 2. Data example

Time	Date	Total
0:00:00	2023/12/10	39
0:15:00	2023/12/10	55
0:30:00	2023/12/10	55
0:45:00	2023/12/10	58
1:00:00	2023/12/10	94

2.2. Method Introduction

ARIMA model, also known as Autoregressive Integrated Moving Average (ARIMA), is a time series prediction method proposed by Box and Jenkins in the early 1970s. It generally uses the weighting of past value, current value and lagging random disturbance term of time series to establish a model, so as to predict the change of future term of time series [11].

ARIMA mainly combines three features for fitting: Autoregressive (AR), Integrated (I) and Moving Average (MA). ARIMA has three different parameters to represent different features, which can help further determine the appropriate values to establish the appropriate model. The parameters are usually expressed as (p, d, q), where p is the coefficient of the autoregressive term, d is the number of differences, and q is the order of the moving average term.

Before using the ARIMA model, the first step is to check whether the time series is stationary. If the time series is not stationary, a differencing is required. Differencing is used to convert a non-time series to a stationary time series. The expression of the first-order differencing formula is:

$$\Delta Yt = Yt - Yt - 1 \tag{1}$$

Where Δ represents the first-order difference operator, Yt represents the observed value of the time series at time t, and $Yt - 1$ represents the observed value of the time series at time t-1. If the D-differencing is required, the first-order differencing formula can be applied repeatedly:

$$\Delta dYt = \Delta d - 1(Yt) - \Delta d - 1(Yt - 1) \tag{2}$$

After obtaining the stationary time series data, the values of p and q are further determined, at which point AR and MA models are needed for confirmation. The AR model assumes that the current value can be expressed as a linear combination of past values, plus an error term. An autoregressive model of order p can be expressed as:

$$Yt = c + \phi 1Yt - 1 + \phi 2Yt - 2 + \cdots + \phi pYt - p + \varepsilon t \quad (3)$$

Where Yt is the observed value of time t and c is the constant term, $\phi 1, \phi 2, \cdots, \phi p$ is the autoregressive coefficient and εt is the error term. The MA model assumes that the current value is a linear combination of the past error plus the current error term. A q order moving average model can be expressed as:

$$Yt = c + \theta 1\varepsilon t - 1 + \theta 2\varepsilon t - 2 + \cdots + \theta q\varepsilon t - q + \varepsilon t \quad (4)$$

Where $\theta 1, \theta 2, \cdots, \theta q$ is the moving average coefficient and εt is the error term. The ARIMA (p, d, q) model is obtained by combining AR, I and MA models.

$$\Delta dYt = c + \phi 1\Delta d - 1Yt - 1 + \phi 2\Delta d - 1Yt - 2 + \cdots + \phi p\Delta d - 1Yt - p + \theta 1\varepsilon t - 1 + \theta 2\varepsilon t - 2 + \cdots + \theta q\varepsilon t - q + \varepsilon t \quad (5)$$

Figure 2 illustrates the work flow of ARIMA.

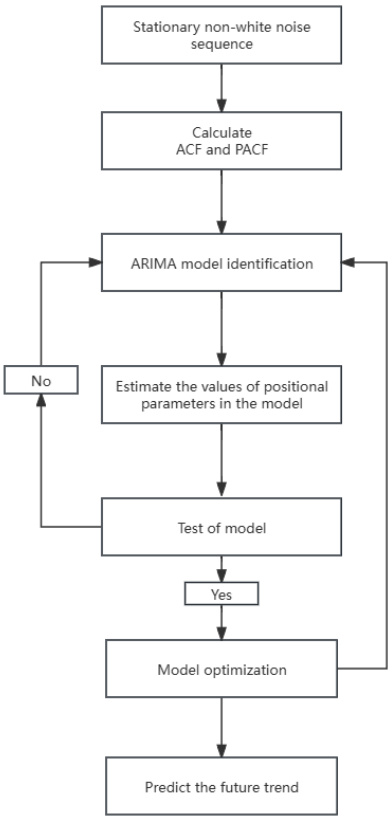


Fig 2. The work flow of ARIMA

3. Results and Discussion

3.1. Time Series Plot

First, the initial data was used to draw the trend chart of the total number of vehicles at the intersection over time, and the data was collected every in 15 minutes, is shown in Figure 3. It can be seen from Figure 4 that the trend of weekly vehicle flow is roughly the same, and the data is not a set of smooth time series data, there are trends or periodic changes.

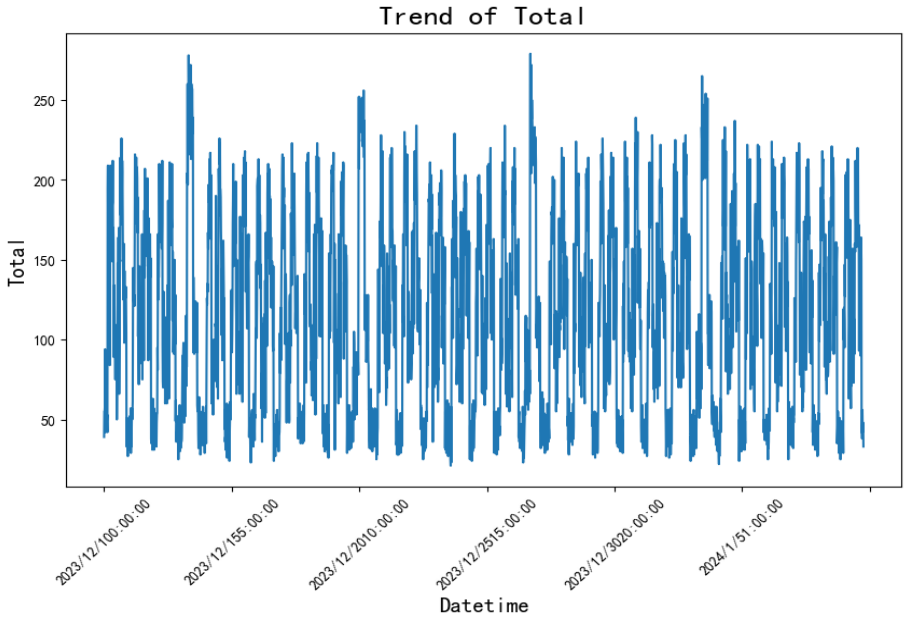


Fig 3. Timing diagram of Total

In order to convert non-stationary time series to stationary time series, first difference is used to make the statistical properties of time series no longer change with time. By subtracting the previous observation from the current observation, horizontal or slope trends in the series can be smoothed out, reflecting the rate of change of the original series and reducing the impact of fluctuations. The result of the first-order difference is shown in Figure 4.

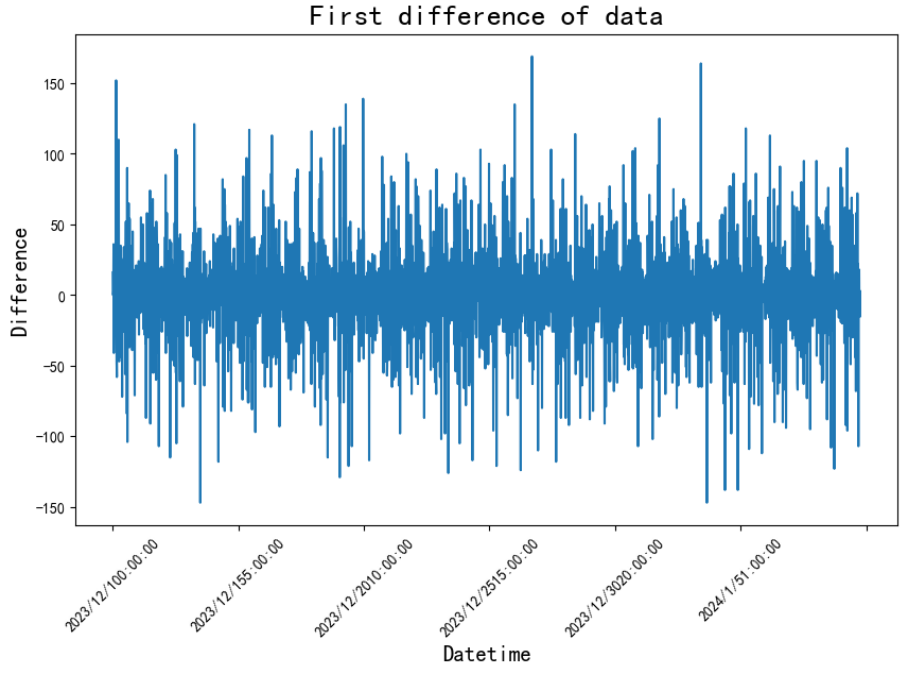


Fig 4. The first difference of data

3.2. ADF Test

In order to further determine whether the first-order difference is stationary, Augmented Dickey-Fuller (ADF) is used to test the stationarity of the data. The null hypothesis of ADF is the existence of unit root. As long as the statistical value is less than 1%, the null hypothesis can be rejected significantly, and the data can be considered stable. The results of ADF are shown in Table 3. T represents the results of ADT, P represents the probability value corresponding to T value, Lag represents the order of lag, 1% represents the critical ADF value of T value under 99% confidence

interval, and 5% represents the critical ADF test value of T value under 95% confidence interval. 10% represents the ADF test value of the T-value at a 90% confidence interval. It can be seen from the table that if the P-value of the first-order difference is less than 0.05 and the T-value is less than 1%, the null hypothesis can be rejected significantly, indicating that the data is stable. Thus, this paper can determine the parameter d=1.

Table 3. Results of ADF test

argument	T	P	Lag	Test	1%	5%	10%
value	-19.502	0	27	2947	-3.433	-2.863	-2.567

3.3. White Noise Test

When the data is pure random data, it is meaningless to analyze the data, so it is necessary to conduct a pure randomness test on the data. In this paper, White Noise (Ljung-Box) is used to test the data, which is used to determine whether there is correlation between any different vectors in the data. White Noise test results are shown in Table 4, where Ib_value represents the test statistic and p_value represents the p statistic based on chi-square distribution. As can be seen from the table, this data rejects the null hypothesis, that is, it is considered that this data is not white noise data and can be used for time series analysis.

Table 4. Partial result of White Noise Test

Ib_value	284.839	286.656	286.657	288.032	289.442	289.995	294.845
p_value	6.6227	5.6668	7.6814	4.1303	1.863	1.1378	7.6287

3.4. PACF and ACF Plot

After the stationarity and white noise are completed, the parameters p and q in the ARIMA model are roughly determined by observing the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) of the time series. ACF plot generally shows a tendency for the correlation between a sequence and its lag value to change as the lag period increases. PACF plot shows the trend of the pure correlation between the sequence and its lag value as the lag period increases. The results of PACF are shown in Figure 5, and the results of ACF are shown in Figure 6. The values of p and q are roughly determined by looking at the trailing and truncated features in the graph. Trailing means that the value is always non-zero, and does not rapidly approach 0 after greater than a certain order, but fluctuates around 0. Truncation is the rapid approach to 0 after a certain order.

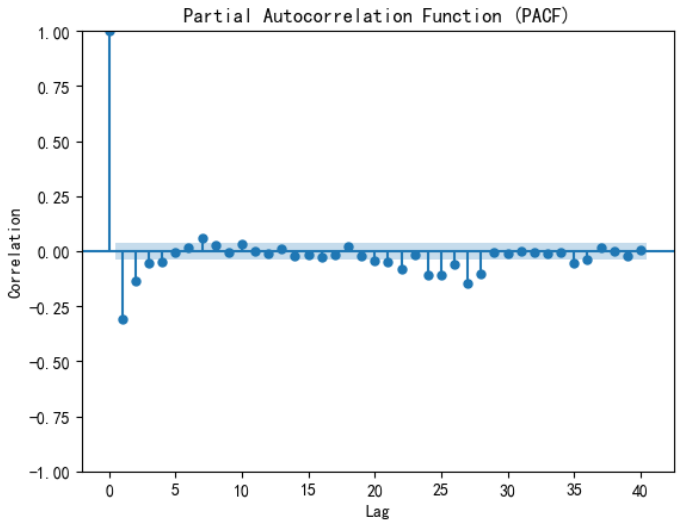


Fig 5. The result of PACF

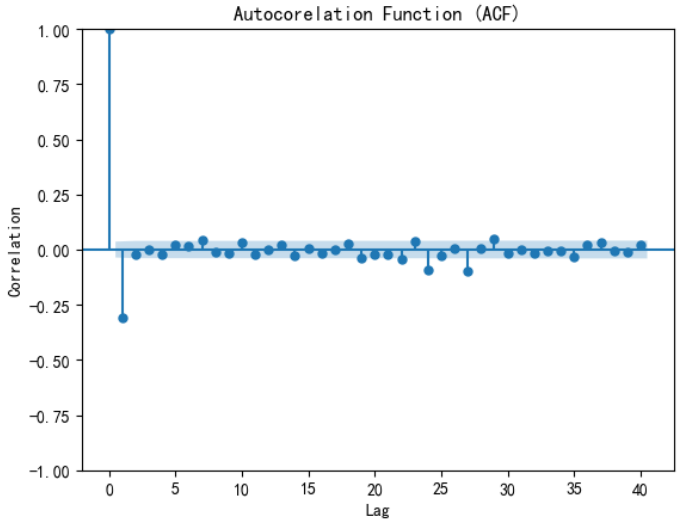


Fig 6. The result of ACF

3.5. Model Results

Bayesian Information Criteria (BIC) criteria can be used to select the order of an ARIMA model. By calculating the BIC values of ARIMA models of different orders, the model with the smallest BIC values can be selected. BIC values usually represent a good balance between model fit and model complexity. The BIC results are shown in Table 5. It can be seen from the table that when p equals 2 and q equals 3, the BIC value is the smallest, which is 27624.25. Thus, the ARIMA model can be determined as ARIMA (2,1,3).

Table 5. Results of BIC

	MA0	MA1	MA2	MA3	MA4
AR0	27963.69	27623.86	27629.30	27637.16	27643.96
AR1	27686.88	27629.40	27637.20	27639.62	27643.88
AR2	27642.24	27637.10	27637.08	27624.25	27631.03
AR3	27639.32	27644.44	27640.62	27648.05	27638.89
AR4	27642.34	27650.21	27633.38	27648.74	27646.86
AR5	27650.17	27653.77	27639.74	27649.32	27665.98

After determining the model, the training was conducted from December 15 to January 4, and the test was conducted from January 5 to January 9, and the line chart of the change of the actual value and the predicted value over time was obtained, as shown in Figure 7. As can be seen from the figure, the predicted value of the ARIMA model fits the actual value to a high degree, and the peak and trough time of the predicted value are roughly the same as the actual value, indicating that the ARIMA model can simulate and predict the traffic flow of urban intersections well.

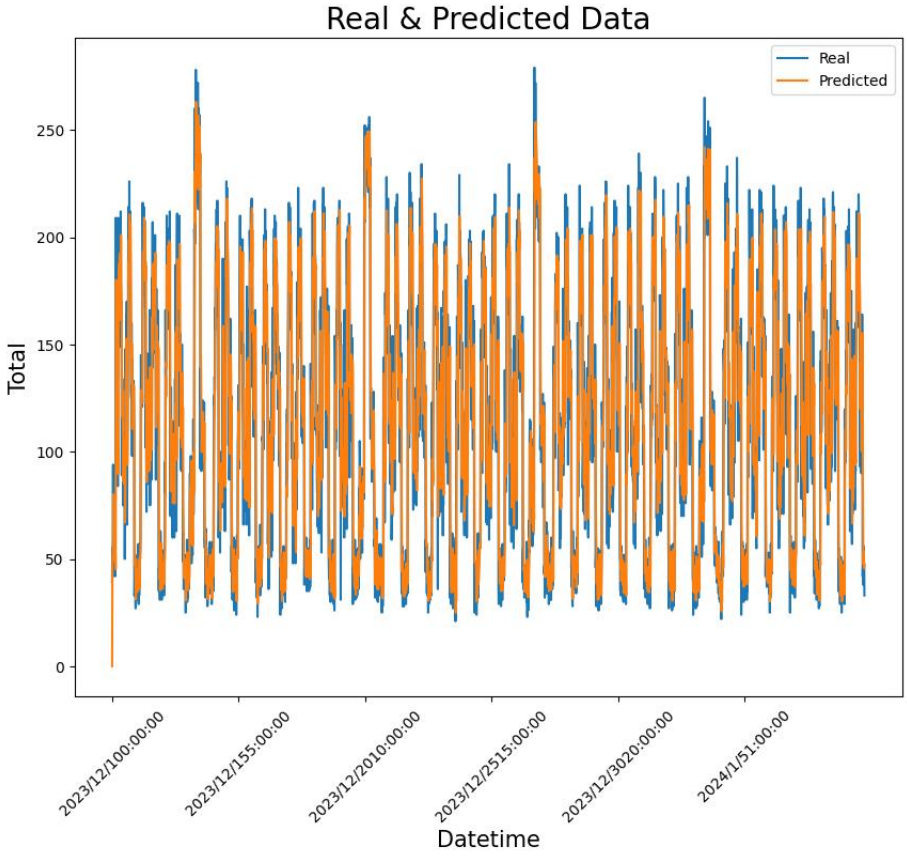


Fig 7. The real data and predicted data

4. Conclusion

This paper aims to use ARIMA model to forecast the traffic flow at urban intersections. By extracting the basic information such as Date, Time, Total of Vehicles and Traffic Situation within a month, the general direction of traffic flow at the intersection is summarized. The required parameters p, d and q are determined according to the trend and characteristics of the data. The data is first converted to stationary time series data by difference, and the value of d is also determined. Then the trailing and truncated features of PACF and ACF were observed, and the p and q values were further determined by combining AR model and MA model. After the selection of parameters, the most suitable ARIMA(p, d, q) model is determined, and the traffic flow is divided into training set and test set according to the requirements, and the time correlation analysis and prediction of the traffic flow are carried out.

ARIMA model is a very classic time series model. it analyzes historical traffic data to predict the traffic flow in the future period of time, and can make efficient and accurate prediction. This prediction can improve the efficiency, safety and sustainability of transportation, thereby contributing to urban development and improving the quality of life of residents.

References

[1] Zhou Jianhe, Liu Ning, Dong Rui, Lu Peng. Traffic Volume Prediction of Expressway Section Based on Algorithm Fusion. Value Engineering, 2024, 43(09): 116-118.

[2] Wang Ting. Research on Highway Traffic Flow Forecasting Method Considering Multiple Factors. Dalian Maritime University, 2022.

[3] Wang Ruidong. Research on Detecting and Forecasting Methods of Urban Road Network Traffic Flow. Xidian University, 2022.

[4] Yan Yufen. Analysis of Intelligent Road Maintenance under Intelligent Transportation. Intelligent Building & Intelligent City, 2024, 5:180-182.

[5] Liu Weiyuan, Ge Yuechun, Li Lei, et al. Research on Holiday Passenger Flow Forecast of Suzhou Rail Transit. Urban Rail Transit, 2021, 34(05): 66-73.

[6] Zhang Guoyun, Jin Hui. Research on Short-Term Passenger Flow Prediction of Urban Rail Transit Based on Improved ARIMA Model. Computer Application and Software, 2022, 39(01): 339-340.

[7] Li Lingling, Xin Hao. Application of time series model ARIMA in data analysis. Fujian Computer, 2024, 40(04): 25-29.

[8] Xiang Junkun, Yu Jiaxing, Gao He, Sun Haoxiang. Constructing SWECPX model based on ARIMA to solve e-commerce demand forecasting problem. China Business Theory, 2024, 8: 29-32.

[9] Fu Lianlian, Fang Qing, Yuan Dongyu, Teng Jiamin. Pig price prediction based on singular spectrum decomposition and LSTM-ARIMA combined model. Journal of Chinese Agricultural Mechanization, 2019, 45(05): 176-181+252.

[10] Lu Youfu, et al. Short-term Traffic Flow Prediction Based on SSA-CNN-BiLSTM combined model. Transportation Research, 2024, 10(01): 18-27.

[11] Li Zheyuan, Zhang Chengpeng. Prediction of the Price Change of Live Pigs in Shandong Province Based on ARIMA Model. Contemporary Rural Finance and Economics, 2024, (05): 34-39.