Research on Traffic Safety Accident Prediction Based on ARIMA

Ziyi Wei*[a], Kefeng Wang[b], Jinhai Gao[c]

[a]School of Energy Science and Engineering, Henan Polytechnic University, Jiaozuo, Henan, 454003, China, 1823601632@qq.com;
[b]School of Energy Science and Engineering, Henan Polytechnic University, Jiaozuo, Henan, 454003, China, kfhere@126.com;
[c]School of Energy Science and Engineering, Henan Polytechnic University, Jiaozuo, Henan, 454003, China, gjh@hpu.edu.cn

Abstract. With the significant increase in the frequency of social traffic activities, traffic safety issues have gradually become the focus of social attention. This paper aims to deeply explore the principles and operating steps of the ARIMA model (autoregressive moving average model) in time series analysis, and evaluate the model's effectiveness in predicting traffic accident data with specific examples. The results show that the ARIMA model performs well in predicting the number of traffic accidents. The error analysis between the predicted data and the actual data provides a basis for further research and verifies the potential application of this method in traffic safety assessment.

Keywords: ARIMA model; Traffic safety accidents; prediction

1 INTRODUCTION

With the continuous improvement of the economic level, traffic congestion and safety problems caused by traffic facilities and equipment have also come along. To deal with this problem, the importance of predicting the number of traffic safety accidents is self-evident. Through scientific and accurate predictions, relevant departments can have a more comprehensive understanding of the changing trends of accident volume and accident rate, and provide data support for optimizing more extensive facilities, planning, and safety management.

In the research on traffic accident prediction, in 2020, Zhang Yifei[1] used the combination of ARIMA and BP neural network, took ships as the research object, and further predicted the accidents of ship traffic in-depth, to guide the decision-making of relevant personnel. In addition, in 2023, Liang Naixing[2] took highway traffic safety as the research object and also used the same ARIMA-LSTM combination method to establish an accident combination prediction model. Not only in China, but also abroad, Nemanja Deretić[3] used the seasonal autoregressive integrated moving average model (SARIMA) to analyze the time series of road traffic accidents. The study found that the time series has obvious seasonal characteristics, and different strategies can be

G. Chen et al. (eds.), *Proceedings of the 2024 International Conference on Rail Transit and Transportation (ICRTT 2024)*, Advances in Engineering Research 254,
https://doi.org/10.2991/978-94-6463-610-9_22

implemented in this prediction context. Praveen Kumar B[4] proposed an optimized ARIMA model, and verified the traffic flow data set completed throughout the day and the traffic flow data set during the morning and evening peak hours through the model. The prediction results show good performance indicators. In short, with the rapid development of my country's transportation industry, traffic safety issues have become complex and changeable. To better provide a reference for traffic safety, this paper will explore the ARIMA forecasting model, and through specific case analysis, compare the forecast results and actual results to verify the accuracy and effectiveness of the proposed forecasting method.

2 ARIMA MODEL

2.1 Prediction Steps of the Model

In 1970, Box and Jenkins proposed the ARIMA model based on the time series analysis model. The ARIMA model assumes that the future forecast value has a solid functional relationship with the current and historical data, that is, it assumes that the future development trend is the same as the current and historical data[5]. The basic steps of the ARIMA model include white noise test, model identification, model fitting of parameter estimation, etc. The specific steps are shown in Figure 1.

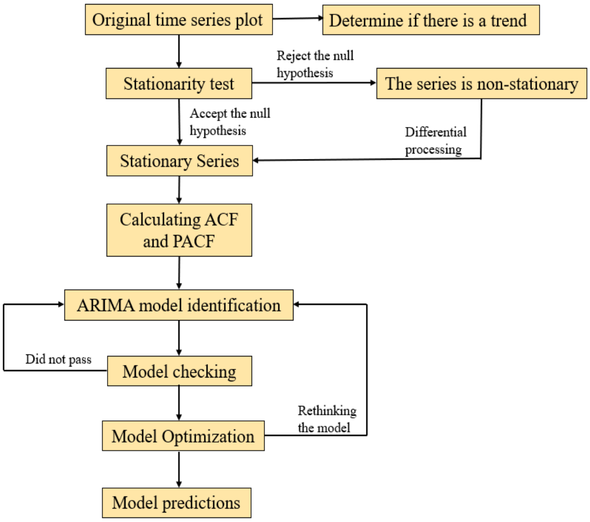

Fig. 1. ARIMA Flowchart.

2.2 Model Formula

In the ARIMA model, AR stands for autoregressive model, I stands for single integral order, and MA stands for moving average model. ARIMA is a combination of the AR model and the MA model. It is determined by three main parameters (p, d, q), and the formula is as follows:

$$y_t = \mu + \sum_{i=1}^{p} \gamma_i y_{t-i} + \varepsilon_t + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} \tag{1}$$

Among them, p represents the order of the autoregressive term, which refers to the number of past values (sequence values lagged p orders) used in the model to predict the current value, that is, the p-order autoregressive model (Auto Regressive Process) formula is expressed as:

$$y_t = \varphi_0 + \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \cdots + \varphi_p y_{t-p} + \varepsilon_t \tag{2}$$

In the formula, φ_i is the lag coefficient, y_{t-i} is the lag term, and ε_t is white noise.

d represents the number of differences, which means that to make the non-stationary time series data stationary, at least d differences are required; q represents the order of the moving average term, also known as the sliding average coefficient, which refers to the number of past error terms used in the model to predict the error term (that is, the error term lags q orders), that is, the q-order moving average model (Moving Average Process) formula is expressed as:

$$y_t = \theta_0 + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q} + \varepsilon_t \tag{3}$$

In the formula, θ_i is the error lag coefficient, ε_{t-i}=F_{t-i}--y_{t-i} is the past forecast error.

3 CASE ANALYSIS

To explore the predictive effect of the ARIMA model on traffic accident data, this paper will use the motor vehicle traffic safety accident data from 2004 to 2022 from the National Bureau of Statistics for predictive analysis.

3.1 Differential Processing

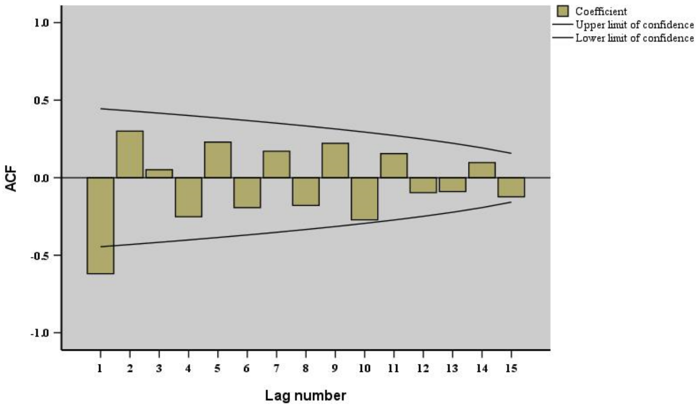

Fig. 2. ACF of 2nd-order difference series.

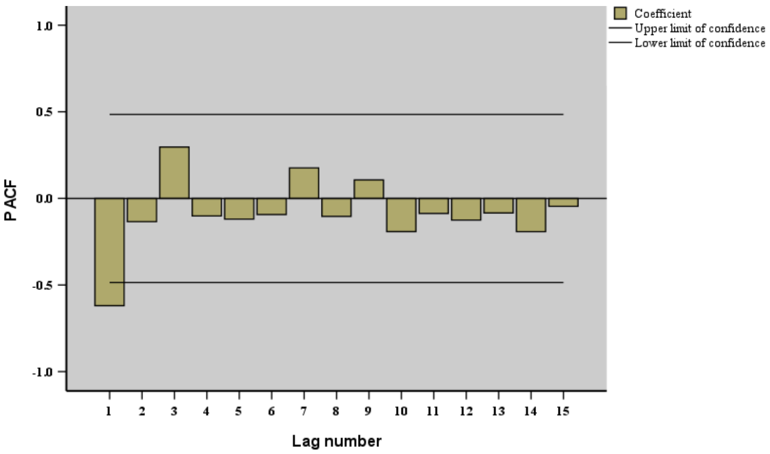

Fig. 3. P ACF of 2nd-order difference series.

The original time series is a non-stationary time series, so this article will perform second-order difference processing on the series. As shown in Figure 2 and Figure 3, the test results of PACF and ACF are first-order tailing, and there is no significant non-zero value after the lag period, indicating that the time series is stationary after the second-order difference. Therefore, it can be determined that the autoregressive order p is equal to 1, the average moving order q is equal to 1, and the difference d is equal to 2.

3.2 Model Building

Based on the values of p, d, and q selected above, an ARIMA (1, 2, 1) model was established in the SPSS system, and the fitting parameters are shown in Table 1. The estimated value of its constant term is 3548.139, that is, the average value of the time series is approximately 3548.139. The parameter of the AR (1) term is 0.686, which represents the strength and direction of the linear relationship between the current observation value and its previous period observation value. The p-value is 0.033, which is lower than 0.05, indicating that the autoregressive term is statistically significant, so the previous period's data has a significant negative impact on the current data. The coefficient of the MA(1) term is 0.188, but its p-value is 0.627, which is much higher than 0.05, indicating that the moving average term is not statistically significant.

Table 1. ARIMA (1, 2, 1) model parameters.

Model	Item	Symbol	Coefficient	Standard error	P-value
ARIMA (1，2，1)	Constant term	c	3548.139	2197.049	0.129
	AR Parameters	α_1	-0.692	0.293	0.033
	MA Parameters	β_1	0.188	0.377	0.627

Based on the above analysis, it is known that MA(1) is not statistically significant, so consider adjusting the model, such as removing insignificant moving average images, to achieve better results.

3.3 Comparison and Adjustment of Models

The model parameters were selected by using the autocorrelation diagram (ACF) and partial autocorrelation diagram (PACF) after differential processing, and the "Expert Modeler" function was not used for modeling. However, to further improve the goodness of fit and prediction accuracy of the model, the SPSS system was used to model the motor vehicle traffic safety accident data from 2004 to 2022 using the "Expert Modeler" while keeping the data set (covering motor vehicle traffic safety accident data from 2004 to 2022) and other external conditions unchanged. At this time, the best model given by the expert modeling is ARIMA (1, 2, 0). Table 2 shows the fitting comparison between ARIMA (1, 2, 1) and ARIMA (1, 2, 0).

Table 2. Comparison table of goodness of fit.

Model	Number of predictors	Model fit statistics				Number of outliers
		R Square	MAPE	RMSE	Normalized BIC	
ARIMA(1, 2, 1)	0	.871	6.521	18606.764	20.163	0
ARIMA(1, 2, 0)	0	.867	6.466	18248.618	19.957	0

As shown in Table 2 neither model uses additional predictors (the number of predictors is 0), which means that they are all based on the historical information of the time series data itself for prediction. R-square is an indicator of the goodness of fit of the model. The closer the value is to 1, the better the model fits. Although the R-square of ARIMA (1, 2, 1) (.871) is slightly higher than the R-square of ARIMA (1, 2, 0) (.867), the difference between the two is very small, indicating that the two models are very close in fitting the data. The mean percentage absolute error (MAPE) and the root mean square error (RMSE) are both indicators of model prediction error. The smaller the value, the more accurate the prediction. Therefore, it is obvious that ARIMA (1, 2, 0) is better than ARIMA (1, 2, 1) in terms of prediction error. Normalized BIC is a criterion that comprehensively considers the goodness of fit and complexity of the model. Lower BIC values usually indicate better models. Here, the normalized BIC value of ARIMA(1,2,0) is lower, indicating that it performs better in terms of model complexity, that is, it has lower model complexity while maintaining similar goodness of fit.

After the above analysis, it is found that the model ARIMA(1,2,0) performs better in terms of goodness of fit, complexity, and prediction error. Therefore, ARIMA(1,2,0) will be selected as the final prediction model in this study.

3.4 Residual White Noise Test

After selecting the ARIMA (1, 2, 0) model, a diagnostic test needs to be performed on it to ensure that the model residuals are white noise sequences. This article will use the autocorrelation function (ACF) and partial autocorrelation function (PACF) graphs to conduct white noise testing.

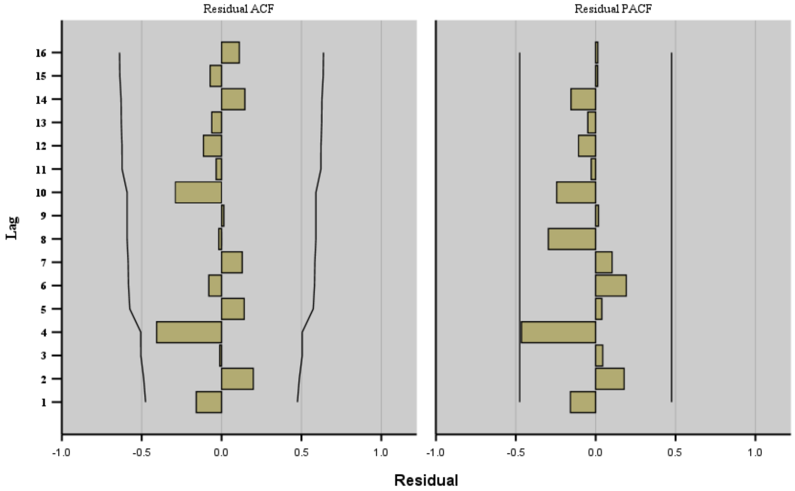

Fig. 4. Fitting residual ACF and PACF plots.

As shown in Figure 4, the ACF and PACF plots show the autocorrelation at different lag stages. After the second-order difference, the autocorrelation coefficients all fall within the confidence interval and approach 0, indicating that the residuals have no obvious deviation and appear as white noise. This shows that the model captures all the information of the data and meets the white noise test.

3.5 Result Analysis

The results of using the ARIMA model to predict the number of traffic accidents are as follows: The curves of the predicted value and the true value are shown in Figure 5. UCL and LCL represent the upper and lower limits of the predicted data, respectively. As the prediction period increases, the upper and lower limits gradually widen, indicating that the uncertainty of the prediction increases. The prediction for 2023 is more accurate, but the prediction uncertainty for 2028 is larger. The MAPE value is 6.466%, that is, the average percentage error is 6.466%. In general, the predicted value of the ARIMA model is within the normal range, and the error is affected by the socio-economic environment.

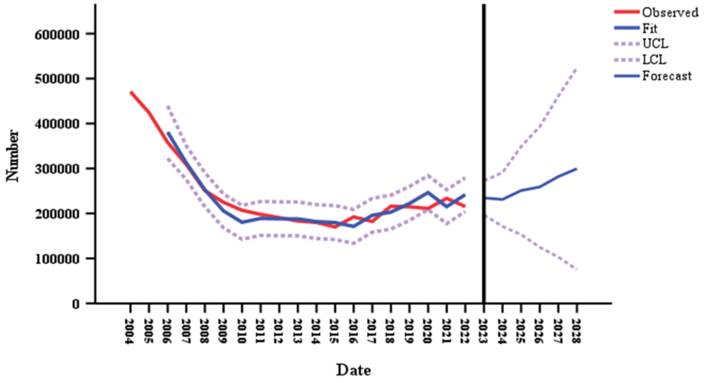

Fig. 5. Fitting result graph.

4 CONCLUSION

This article examines traffic accident prediction using the ARIMA model and reaches the following conclusions: 1)The ARIMA model effectively forecasts future traffic accidents, enabling management departments to assess current traffic safety and adjust measures accordingly; 2)Despite its effectiveness, the ARIMA model has limitations, such as not accounting for the impact of emergencies like extreme weather. Future research should consider integrating advanced models, such as BP neural networks and XGBoost, to enhance prediction accuracy.

ACKNOWLEDGMENTS

Project name: Soft Science Research Project of Henan Provincial Department of Science and Technology, Project No.242400410292. Project name: The Fundamental Research Funds for the Universities of Henan Province, Project No.NSFRF230425.

REFERENCES

1. Zhang, Y. and Fu, Y., Ship traffic accident prediction based on ARIMA-BP neural network, Journal of Shanghai Maritime University, 47-52 (2020).
2. Liang, N., Yuan, J. and Yang, W., et al., Research on highway traffic safety combination prediction model based on ARIMA-LSTM, Journal of Chongqing Jiaotong University (Natural Science Edition), 131-138(2023).
3. Deretić, N., Stanimirović, D. and Awadh, M. N., et al. SARIMA modelling approach for forecasting of traffic accidents, 14(8): 4403(2022).
4. Kumar, P. B. and Hariharan, K., Time series traffic flow prediction with hyper-parameter optimized ARIMA models for intelligent transportation system, Journal of Scientific & Industrial Research, 408-415(2022).
5. Zhang, F., Comparison of traffic flow prediction methods for urban transportation networks, Science and Technology Progress and Countermeasures, 57-59(2004).