Short-term Traffic Flow Prediction for Expressway based on ARIMA: Compared with LSTM

Jinghan Zou

College of Computer and Cyber Security, Fujian Normal University, Fuzhou, 350000, China
u2389788@unimail.hud.ac.uk

Abstract. Predicting short-term traffic flow holds significant value for the operation and management of highway traffic. As a result, it's crucial to develop a feasible short-term traffic flow forecast model and use traffic flow data for prediction in an efficient manner. In this study, for predicting traffic flow, the Autoregressive Integrated Moving Average (ARIMA) model is employed. Firstly, the original sequence of highway traffic flow data is differenced and the ADF test is used to verify the stationarity of the sequence. Subsequently, the value range of p and q is preliminarily determined by the graphical feature analysis method, and then the model is constructed and evaluated according to the AIC and BIC criteria to determine the values of p and q. Next, to estimate parameters, utilize maximum likelihood estimation and construct a complete ARIMA model. After the model is established, a white noise test is performed on the residuals to ensure a high degree of model fit. Finally, a static forecasting method is employed to predict real traffic flow data from Luxembourg, and the outcomes of the model's predictions are assessed. The experimental findings show that the ARIMA model predicts short-term highway traffic flow with great accuracy and dependability.

Keywords: Highway traffic, Short-term traffic flow, ARIMA model, prediction, LSTM model.

1 Introduction

A well-developed road network is an essential component of the national economy and can promote regional economic development. In recent years, with the improvement of people's living standards and the continuous acceleration of urbanization, the number of cars owned by people has continued to increase. Expressways, due to their traits of high speed, large capacity, and comfort, have become the preferred route for more and more vehicles, playing a vital role in road network transportation. However, the surge in traffic volume has also caused frequent traffic congestion problems, which not only affect the travel experience but also pose a challenge to road safety [1]. Although current detection technology can monitor highway traffic conditions in real time, travelers and traffic management departments are more concerned about traffic conditions in the short term. Accurate short-term traffic flow forecasts can not only

K. Subramanian (ed.), *Proceedings of the International Workshop on Advances in Deep Learning for Image Analysis and Computer Vision (IWADIC 2025)*, Advances in Computer Science Research 128,
https://doi.org/10.2991/978-94-6239-648-7_9

assist traffic management departments in optimizing their decisions, but also provide travelers with effective route guidance, thereby improving overall traffic efficiency [2].

Short-term traffic volume forecasting is to analyze and forecast time series based on a few minutes as a data unit, and to predict the traffic volume in the short term in the future in real time. It works well in situations requiring real-time dynamic traffic control and guiding [3]. In this study, traffic flow prediction methods are currently divided into four categories according to the different models used [4]: The first category is linear theoretical models, such as Wang et al. (2003) using the extended Kalman filter model to design a traffic state estimator (EKF) to achieve dynamic prediction of conventional traffic variables [5]. The second type of nonlinear theoretical model, such as Jun et al. (2008), used artificial neural networks (ANN) to achieve traffic flow prediction, demonstrating the adaptability of nonlinear models [6]. The third type of combination model, such as Ma et al., combined the autoregressive integrated moving average model (ARIMA), grey wolf optimizer (GWO) and long short-term memory network (LSTM), and proposed a method based on the ARIMA-GWO-LSTM combination model, which effectively improved the accuracy of highway traffic flow prediction [7]. The fourth type of other models, such as Wang et al., suggested a model for dynamic traffic prediction that realizes the dynamic evolution prediction of traffic status through real-time processing of traffic flow data [8].

To improve the accuracy of short-term freeway traffic flow prediction and optimize traffic management efficiency, this study uses the ARIMA model to predict traffic flow. This research establishes an ARIMA-based short-term freeway traffic flow prediction model and uses measured data from the Luxembourg Highway for a case study.

2 Method

2.1 Data Source and Description

This study's data comes from the " Motorway Traffic in Luxembourg Dataset ". The original data was recorded from November 19 to December 26, 2019, covering 186 monitoring points along seven major highways in Luxembourg. The data acquisition system continuously records the traffic flow through each monitoring section at 5-minute intervals, and the total of raw data is approximately 4.2 million points. To focus on the research objectives, this study only retains a complete 72-hour data record from November 20 to 22, 2019, for a single direction on the A1 highway, according to Figure 1.

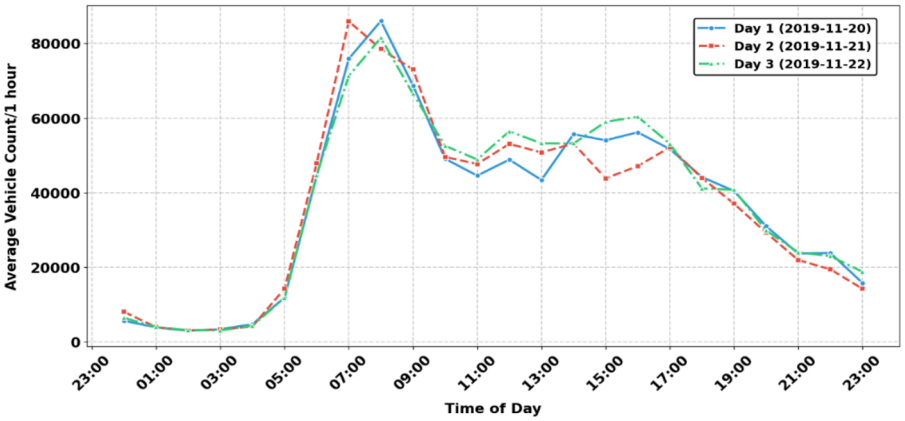

Fig. 1. Vehicle volume from November 20th to 22nd (Data from: Motorway Traffic in Luxembourg Dataset).

2.2 Indicator Selection and Description

In this study, in order to focus on exploring the temporal variations in passenger flow better, firstly filter the raw data to identify the desired categories and process them. The results are shown in Table 1. This study selects the five-minute passenger flow on Expressway A1 as the core indicator. This is because the temporal variation in expressway traffic flow provides a direct reflection of the fluctuating characteristics of traffic load on a given road section allowing management to dynamically adjust traffic control measures, thereby improving road network efficiency.

Table 1. Processed data.

Column names	Type	Instruction	Value range
timestamp	String	Time period	[2019.11.20-2019.11.22]
total_vehicle_count	Integer	Total traffic volume	[1636- 110106]

2.3 Method Introduction

The ARIMA model is a statistical model proposed by Box and Jenkins in 1970 and frequently employed in time series forecasting. The essence of this model is to stabilize the data and use the autoregressive moving average model (ARMA) for linear fitting [9]. The model is abbreviated as ARIMA (p, d, q), and its core structure is as follows：

The ARMA model is an important time series analysis model, which essentially blends the autoregressive model (AR) and the moving average model (MA)[9].

As an important part of the ARMA model, the core principle of the AR model is to use the linear combination of historical observations to build a forecasting model. The model order p is a key parameter, which represents the number of historical observations taken into account when building the model. The following is the mathematical expression：

$$x_t = a + \varphi_1 x_{t-1} + \varphi_2 x_{t-2} + \cdots + \varphi_p x_{t-p} + \epsilon_t \quad (1)$$

In the formula：a is a constant，φ_p is the autoregressive coefficient；ϵ_t is the random error term, just known as the white noise sequence value；x_t x is the time series value at time t; p is the autoregressive period, also known as the order of the AR model. The MA model is a great method for time series forecasting and is constructed based on a linear combination of historical forecast errors. In this model, the order q represents the number of error terms. The following is the mathematical expression：

$$x_t = a + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots + \theta_q \epsilon_{t-q} + \epsilon_t \quad (2)$$

In the formula: θ_q is the moving average coefficient;ϵ_t is the white noise error term of the time series; q is the moving average period, also known as the order of the MA model. Therefore, the ARMA(p,q) expression is:

$$x_t = c + \beta_1 x_{t-1} + \beta_2 x_{t-2} + \cdots + \beta_p x_{t-p} + \epsilon_t - \theta_1 \epsilon_{t-1} - \cdots - \theta_q \epsilon_{t-q} \quad (3)$$

Difference operation: Perform a difference operation on a non-stationary sequence to achieve a stable state. The parameter d represents the number of differences. The difference formula is:

$$(1 - B)x_t = \nabla^d x_t = \nabla^{d-1} x_t - \nabla^{d-1} x_{t-1} \quad (4)$$

The complete ARIMA(p,d,q) model can be expressed as:

$$x_t = c + \varphi_1 x_{t-1} + \varphi_2 x_{t-2} + \cdots + \varphi_p x_{t-p} - \theta_1 \epsilon_{t-1} - \theta_2 \epsilon_{t-2} - \cdots - \theta_q \epsilon_{t-q} \quad (5)$$

3 Results and Discussion

3.1 ADF Test

Determining the stationarity of a series is a key step in the ARIMA modeling process. Figure 1 shows that traffic flow increases significantly after 5:00 AM and decreases significantly after 5:00 PM. The data fluctuates significantly, exhibiting distinct time-of-day characteristics and trend changes, thereby preliminarily determining that this time series data is non-stationary. To this end, a first-order differencing operation is performed to eliminate the trend and make the data stationary. Then use the ADF test to verify the stationarity of both the original series and the first-order difference series. Table 2 displays the findings.

Table 2. ADF Test.

Difference order	t	p	1% Level	5% Level	10% Level
0	-2.201	0.205	-3.43	-2.86	-2.56
1	-3.136	0.023	-3.45	-2.87	-2.57

Initially, assume that the original time series is stationary. However, Table 2 presents that the P value of the original flow series exceeds 0.05 and the t-statistic is greater than the 10% confidence level critical value, so the null hypothesis should be rejected. Then assume that the first-order difference series is stationary. The ADF test shows that its P value is much less than 0.05 and the t-statistic is less than the 5% confidence level critical value, confirming that the first-order difference series is stationary, with no random walk, long-term trend, or seasonal changes. Therefore, subsequent research is based on first-order difference series. Figure 2 further confirms that the time series data after the first-order difference is stable, showing random fluctuations and no significant trend. Based on this, when constructing the ARIMA(p, d, q) model, the d value is set to 1, i.e., using ARIMA(p, 1, q) for short-term traffic passenger flow forecasting and analysis.

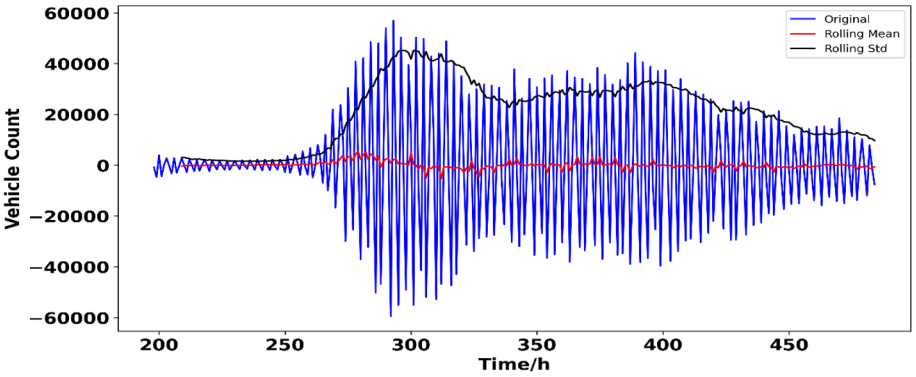

Fig. 2. Plot of traffic volume with first-order differences (Picture credit: Original).

3.2 Model Identification

This study uses both the graphical feature recognition method and the AIC and BIC criteria to determine the order of the model. Firstly, after determining the difference order d=1, the order of the model is preliminarily established by the characteristics of the autocorrelation plot (ACF) and partial autocorrelation plot (PACF) of the sequence data. As shown in Figure 3, the autocorrelation coefficient gradually decreases after the fifth order and never fully falls within the confidence interval, so it can be determined that there is tailing after the fifth order. The partial autocorrelation coefficient in Figure 4 also generally falls within two standard deviations after the fifth order, indicating that there is tailing after the fifth order. Therefore, preliminarily determine that both p and q of the ARIMA (p, 1, q) are less than 5.

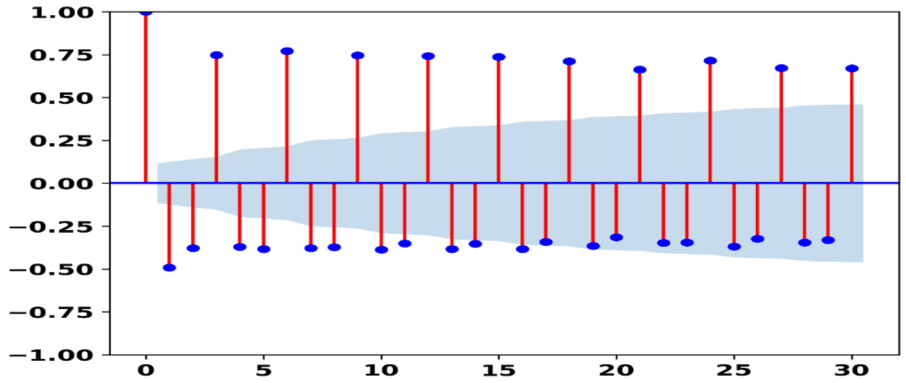

Fig. 3. ACF Plot(Picture credit: Original).

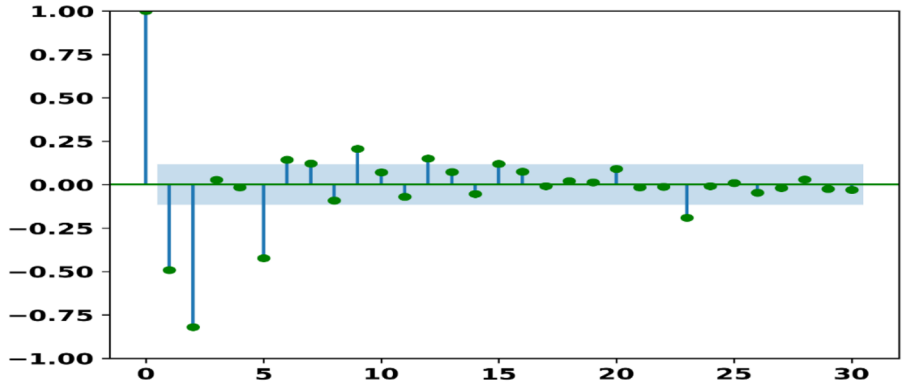

Fig. 4. PACF Plot(Picture credit: Original).

Based on the preliminary observation results, models with different (p, q) order combinations are constructed and evaluated according to the AIC and BIC criteria. The optimal model combination is finally determined and listed in Table 3 (only some order combination results are shown). According to the data, by comparing and analyzing its AIC, BIC and log-likelihood values, it can be seen that the log-likelihood value of ARIMA (3, 1, 4) is relatively large, and the AIC and BIC values are the smallest, so the final model order is determined to be p = 3, q = 4, d = 1, that is, ARIMA (3, 1, 4). Furthermore, using maximum likelihood estimation for parameter estimation, the complete ARIMA (3,1,4) model is obtained, which can be expressed as:

$$x_t = -0.272x_{t-1} + 0.011x_{t-2} + 0.985x_{t-3} - 0.860x_{t-4} + \epsilon_t - 0.730\epsilon_{t-1} + 0.02\epsilon_{t-2} - 0.695\epsilon_{t-3} + 0.604\epsilon_{t-4} \quad (6)$$

Table 3. Evaluation Test.

Rank	Model	AIC	BIC	Log-Likelihood
1	ARIMA(3,1,4)	6120.34	6149.62	-3052.2
2	ARIMA(4,1,4)	6124.89	6157.82	-3053.4
3	ARIMA(2,1,3)	6128.73	6150.69	-3058.4
4	ARIMA(2,1,4)	6130.23	6155.85	-3058.1
5	ARIMA(3,1,3)	6130.23	6155.85	-3058.1
6	ARIMA(4,1,3)	6131.89	6161.16	-3057.9
7	ARIMA(4,1,2)	6184.34	6209.96	-3085.2
8	ARIMA(4,1,1)	6197.15	6219.11	-3092.6
9	ARIMA(2,1,0)	6198.39	6209.37	-3096.2
10	ARIMA(3,1,0)	6200.19	6214.83	-3096.1

3.3 Model Testing

After establishing the model, a white noise test needs to be performed on the residuals to assess the model's goodness of fit. If the residuals conform to a white noise sequence, the model fit is relatively effective. As shown in Figure 5, in general, the fitted values' changes align with the actual values, presumably indicating that the model's residual sequence is stationary. Therefore, a Ljung-Box test was performed on the residuals of the ARIMA(3,1,4) model. The result shows a p-value of 0.88, significantly greater than 0.05, indicating that the significance test is passed. This indicates that the ARIMA(3,1,4) model has an excellent fit.

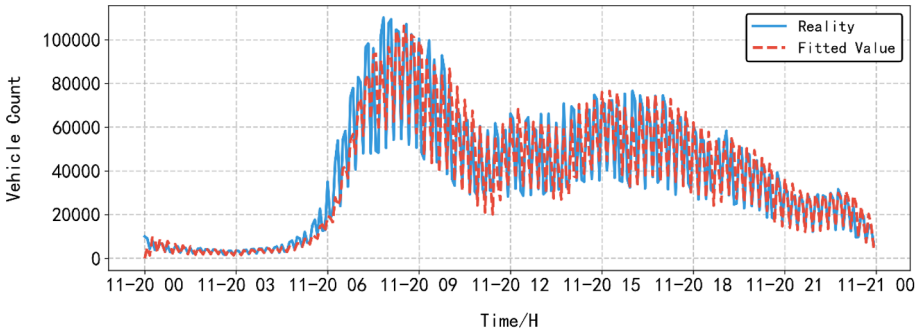

Fig. 5. Traffic Fitting Diagram on November 20(Picture credit: Original).

3.4 Model Prediction

The traffic flow data on January 20, 2019, is used to obtain the prediction model ARIMA (3, 1, 4), as shown in Equation 6. 573 traffic flow data items are predicted for the two working days of January 21 and 22. The prediction results are shown in Figures 6 and Figure 7.

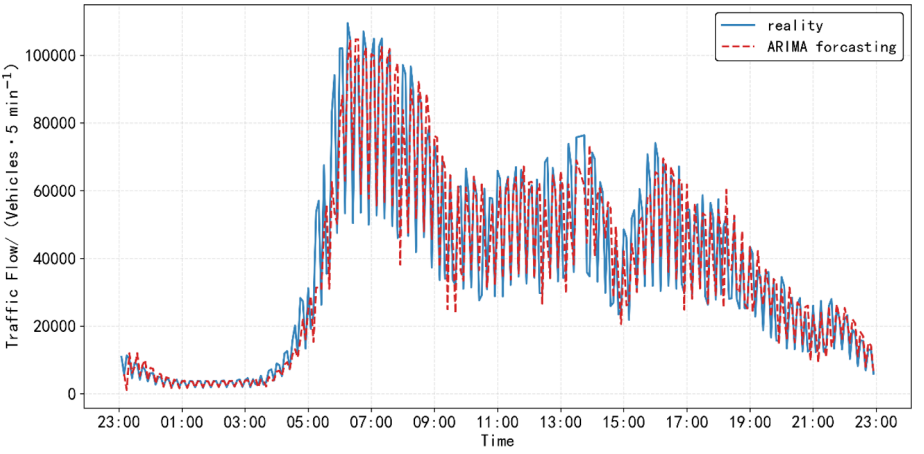

Fig. 6. Traffic Forecast for November 21st(Picture credit: Original).

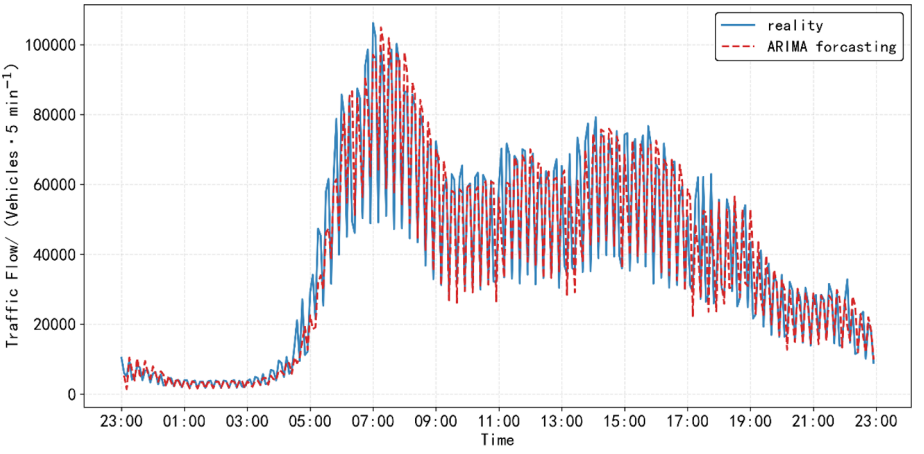

Fig. 7. Traffic Forecast for November 22nd(Picture credit: Original).

From the comparison chart of actual and predicted values, it can be seen that the model performs well in predicting short-term traffic volume data. Specifically, the curves for the predicted and true values closely match, demonstrating a certain degree of feasibility. However, due to the high level of traffic volume and uncertainty between 5:00 AM and 6:00 PM, the prediction performance for this period is relatively poor. But overall the prediction performance for the entire time period is good, demonstrating that the model can predict future traffic volumes to a certain extent and demonstrates good performance in short-term traffic volume forecasting.

3.5 Comparative Analysis of Forecasting Methods

The LSTM model is a special kind of recurrent neural network (RNN) that can learn long-term dependencies and retain long-term historical information, making it powerful in processing time series and very suitable for traffic flow prediction [10]. The prediction results of the LSTM model and the model in this study on the same data set are compared and analyzed. In terms of prediction accuracy, MAE, MAPE, and RMSE are used to analyze the prediction results. The analysis results are shown in Tables 4 and 5 respectively.

Through comparative analysis, it is evident that the ARIMA model's MAE, RMSE, and MAPE are all smaller than those of the LSTM model, indicating that the ARIMA model has a good advantage in processing traffic flow data with periodicity and trends, and shows good prediction accuracy and precision.

Table 4. ARIMA Prediction Results Analysis.

	MAE	RMSE	MAPE
21 day	6025.7	9976.6	18.86%
22 day	4891.8	7829.4	16.25%

Table 5. LSTM Prediction Results Analysis.

	MAE	RMSE	MAPE
21 day	12225.1	15405.5	44.79%
22 day	12100.5	14853.6	44.88%

4 Conclusion

This study uses short-term traffic flow data from the 20th as a training set for model calibration. Through ADF stationarity testing and stabilization, model identification, parameter estimation, and residual testing, an ARIMA (3, 1, 4) forecasting model is established. Then a static forecasting method is used to forecast traffic flow data from the 21st to the 22nd, and the results are compared with those from the LSTM forecast. The analysis results demonstrate that this model presents good real-time responsiveness and feasibility in short-term traffic volume forecasting. In future research, external uncertainties such as climate change and sudden accidents can be introduced, and the ARIMA model can be integrated with nonlinear models (like LSTM) for prediction to enhance the prediction defect of the ARIMA model that does not consider the interference of uncertain factors, in order to boost the model's accuracy and the prediction's precision.

References

1. Li, D.-m., Liu, B.: Modeling and Prediction of Highway Traffic Flow Based on Wavelet Neural Network. In: Proceedings of the 2014 International Conference on Machine Learning and Cybernetics, pp. 675–679. IEEE, Lanzhou, China (2014)
2. Tang, Y., Liu, W., Sun, D., et al.: Application of Improved Time Series Model in Short-Term Highway Traffic Flow Prediction. Journal of Computer Applications Research 32(1), 146–149 (2015)
3. Zhang, T., Yuan, P.: Short-Term Traffic Volume Prediction Model Based on ARIMA. Intelligent Computer and Applications 10(7), 273–278 (2020)
4. Xu, H., Chen, G., Wang, S.: Short-Term Traffic Flow Prediction Based on RNN Model. Today's Manufacturing and Upgrading, 09, 50–52 (2023)
5. Wang, Y., Papageorgiou, M., Messmer, A.: Motorway Traffic State Estimation Based on Extended Kalman Filter. In: Proceedings of the 2003 European Control Conference (ECC), pp. 1934–1939. IEEE, Cambridge, UK (2003)
6. Jun, M., Ying, M.: Research of Traffic Flow Forecasting Based on Neural Network. In: Proceedings of 2008 Second International Symposium on Intelligent Information Technology Application, pp. 104–108. IEEE, Shanghai, China (2008)
7. Ma, C., Gu, K., Zhao, Y., et al.: Research on Highway Traffic Flow Prediction Based on a Hybrid Model of ARIMA-GWO-LSTM. Neural Computing and Applications 37, 14703–14722 (2025)
8. Wang, Y., Chen, Y., Qin, M., et al.: Dynamic Traffic Prediction Based on Traffic Flow Mining. In: Proceedings of the 2006 6th World Congress on Intelligent Control and Automation, pp. 6078–6081. IEEE, Dalian, China (2006)
9. Cui, J., Li, Z., Zhao, J., et al.: Short-Term Highway Traffic Flow Prediction Method Based on ARIMA. Journal of Transportation Engineering 23(4), 145–154 (2023)
10. Poonia, P., Jain, V. K.: Short-Term Traffic Flow Prediction Using LSTM. In: Proceedings of the 2020 International Conference on Emerging Trends in Communication, Control and Computing (ICONC3), pp. 1–4. IEEE, Lakshmangarh, India (2020).