Road Condition Prediction Based on ARIMA Algorithm

Yiwei Wang * Yuhao Zhang, Yanbing Zhou and Wenjin Tang

Navigation College, Dalian Maritime University, Dalian, China

* Corresponding author: 790182948@qq.com

Abstract. This study investigates a road condition prediction method based on autoregressive integral moving average model (ARIMA) and its application in road traffic management and planning. With the acceleration of urbanization and the increasing demand for transportation, effective road condition prediction is crucial for traffic management and planning. This paper first introduces the challenges faced in the field of road transportation. Then, the principles and applications of the ARIMA algorithm in road condition prediction are elaborated, focusing on its advantages in capturing trends and cyclical changes in road traffic data. Subsequently, this paper verifies the effectiveness and usefulness of the ARIMA algorithm in road condition prediction through empirical analysis and case studies. The results show that the ARIMA algorithm exhibits high accuracy and stability in short- and medium-term road condition prediction, providing a simple and effective prediction tool for traffic management authorities and planners. Finally, this paper provides an outlook on the future research direction, presenting research outlooks on model optimization and improvement, combining other methods, and real-time prediction and application, in order to further improve the accuracy and practicality of road condition prediction.

Keywords: Road condition, ARIMA, Prediction.

1. Introduction

As an indispensable part of modern urban life, road transportation bears the important responsibility of connecting all corners of the city, promoting economic development and improving people's lives. However, with the continuous acceleration of urbanization, the surge in the number of private cars and the complexity of urban planning, road traffic congestion and unpredictable changes in road conditions have become common challenges in daily life. These problems have a direct impact on transportation efficiency, environmental quality, and the travel experience of urban residents. Therefore, accurate prediction and effective management of road traffic conditions have become crucial.

In this paper, road traffic prediction has become a research topic of great interest in the field of transportation management and planning. Accurate road condition prediction can help traffic management departments to optimize traffic signal control and improve road planning, as well as provide smarter travel suggestions for drivers, thus improving overall traffic efficiency and reducing traffic congestion. Therefore, the development of effective road condition prediction models is important for the realization of intelligent transportation systems and sustainable urban development.

In the past decades, researchers and experts have proposed many road traffic prediction models, including various approaches based on statistical methods, machine learning techniques, and time series analysis. Among them, Autoregressive Integral Moving Average (ARIMA) model, as a classical time series analysis method, has attracted much attention due to its superior performance in modeling and forecasting time series data. ARIMA model has good interpretability and modeling flexibility for various types of time series data, and thus has a broad application prospect in the field of road traffic forecasting.

Chen et al [1] proposed an ARIMA - LR hybrid model based on autoregressive summation moving average model (ARIMA) and linear regression model (LR) using an improved Bayesian combination model. It is found that the hybrid ARIMA - LR model not only better adapts to the changes brought about by unexpected events, but also has a higher overall prediction accuracy than the ARIMA model and the LR model through the prediction of the actual civil aviation cargo volume. The average absolute error (MAE), mean square error (MSE) and average absolute percentage error (MAPE) of

the hybrid ARIMA - LR model are lower than those of the ARIMA model by 1.06, 29.02, and 0.03, respectively, and compared to the LR model, they are lower by 3.00, 92.00, and 0.06, respectively. Rezaiy et al [2] proposed a method that combines the wavelet transform with an autoregressive integrated moving average (ARIMA) model, a novel hybrid model called wavelet ARIMA (W-ARIMA), to improve the accuracy of drought prediction. The paper carefully analyzes monthly precipitation data from January 1970 to December 2019 in Kabul, Afghanistan, focusing on multiple time scales (SPI 3, SPI 6, SPI 9, SPI 12). Comparison with the conventional ARIMA method demonstrates the superior performance of the proposed W-ARIMA model. Key statistical metrics, including root mean square error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE), emphasize the progress made by the W-ARIMA model, especially in SPI 12 prediction. Jin et al [3] conducted a study on the performance of the three combined models, namely PSO-LSTM-ARIMA, MLR-LSTM-ARIMA, and BPNN-LSTM- ARIMA to compare the MSE, RMSE and MAE of three combination models. It was found that the third model containing MSE, RMSE and MAE has the best prediction accuracy. The study selected LSTM model and ARIMA model. First, it used a single model to predict the German pandemic data. It then merged them using BP neural networks, particle swarm methods and multiple linear regression. To confirm this finding, the paper re-predicted the Japanese pandemic data and retrieved the MSE, RMSE and MAE values of the BPNN-LSTM-ARIMA model as 6141895.956, 2478.285 and 1249.832, respectively. the most accurate model was still this combined model. The study shows that BP neural network coupled with LSTM model and ARIMA model has the highest prediction accuracy.

For traffic flow prediction, Hu et al [4] proposed a traffic flow prediction method based on dynamic multi-graph convolutional network (GTDMGCN). Instead of a single graph, the graph transformation network proposed in this paper constructs multiple graphs to modulate a complex traffic network. Based on this, a time-domain gate convolution method is proposed to obtain the time-domain characteristics of traffic flow. The proposed GTDMGCN model is evaluated on four real traffic datasets, PEMS03, PEMS04, PEMS07, and PEMS08, and the average increments of MAE, RMSE, and MAPE metrics are 9.78%, 7.80%, and 5.96%, respectively, compared to the existing results. Xia et al [5] proposed a new dynamic graph-based deep learning framework and a dynamic graph recurrent network for traffic flow prediction called dynamic spatio-temporal graph recurrent neural network. In this framework, a new dynamic graph generator is designed to obtain dynamic representations of nodes, which employs a multi-head attention network and dynamic node embedding to capture hidden spatial dependencies more efficiently. In order to infer the edge states of the dynamic graph at different moments, the generated dynamic graph is trained as special time-series data for downstream time-series prediction via a dynamic graph recurrent neural network. In contrast to methods that directly connect static and dynamic graphs, a new fusion framework integrates a two-channel convolutional network with a penalty term and a gate fusion layer to extract dynamic spatial dependencies from multiple graphs to improve prediction accuracy and reduce computational consumption. Experiments are conducted on three real datasets to evaluate the superior performance of the proposed model. The performance of the proposed method on the three datasets is improved by 10%-26% compared to the previous state-of-the-art baseline.

The aim of this thesis is to explore a road condition prediction method based on the ARIMA algorithm for analyzing and modeling time series data with a view to providing more accurate predictive information for road traffic management and planning. This study will explore in depth the principle of ARIMA algorithm, its specific application in road condition prediction, in order to verify the effectiveness and practicality of ARIMA algorithm in road condition prediction. Through the in-depth discussion in this study, we hope to provide new ideas and methods to improve urban road traffic management and planning, and contribute to the realization of intelligent transportation and sustainable development.

2. Data Smoothness

There are many measures of road conditions, such as congestion distance, total number of congested vehicles, and so on. In this work, the average speed of a representative congested roadway is chosen to be measured. Crawl a certain period of continuous time under the average speed of a roadway information. Line graph shown in Figure 1, the regression model (AR) requires the data to have smoothness, smoothness describes the curve in the future period of time can still follow the existing pattern of "inertia" continues, that is, the mean and variance of the series does not change significantly.

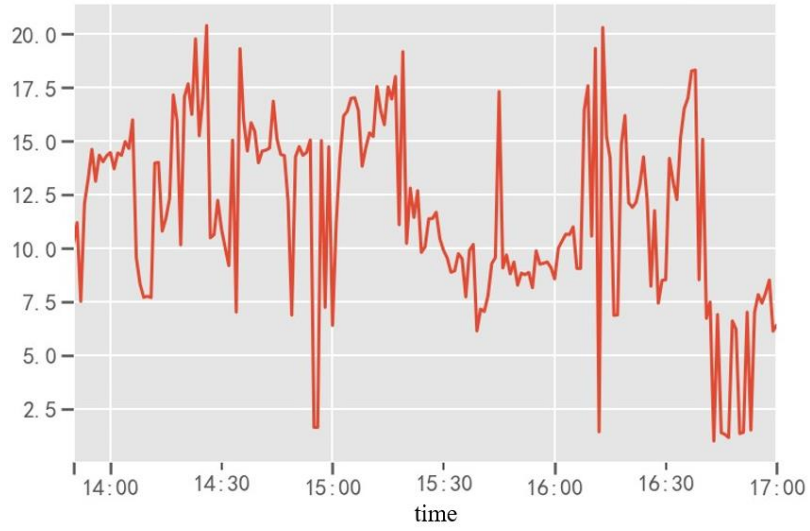

Figure 1. Sequence diagram of the original data

Smoothness in ARIMA is categorized into strictly smooth and weakly smooth, and the data we crawled more or less changes over time, making it difficult to have a normal distribution, so the default data set is weakly smooth. This requires that the expectation and correlation coefficients of the data remain unchanged or float in a small range. Because the value Xt of t at some point in the future depends on its past information, if the dependency changes too much, it cannot be predicted.

The idea of the difference method is to use the difference between the time series at t and t-1 as the target data, which is smoother. The data obtained by differencing can also continue to be differenced, the number of times the difference is the order of that difference d, taking that order will be used in the determination of parameter d. As Figure 2 shows the first order difference, the smoothness of the difference result can be observed, so it is very easy to determine the order of the difference. The data after the first order differencing is clearly smoother than the original data.

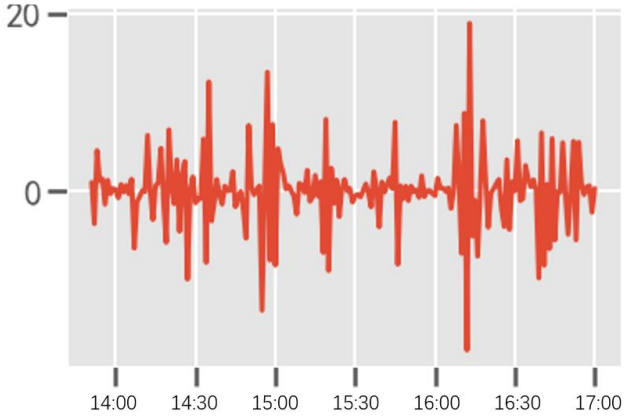

Figure 2. Line graph of first-order difference data

3. ARIMA model

3.1. Autoregressive model (AR)

Autoregression describes the relationship between current and historical values, using the variable's own historical time data to make predictions about itself. The autoregressive model must satisfy the smoothness requirement, and the equation for the pth order autoregressive process defines.

$$y_t = \mu + \sum_{i=1}^{p} r_i y_{t-i} + \varepsilon_t \quad (1)$$

In this formula, yt is the current value, μ is the constant term, p is the order, ri is the autocorrelation coefficient, and εt is the error.

Where p describes that the current value is related to the values at the previous p times, which need to be specified, and the coefficient ri can be solved by a variety of methods such as maximum likelihood estimation, least squares method, and so on.

Autoregression is only applicable to predict the phenomenon related to its own prior period, and is applicable to the prediction of future road conditions in this work.

3.2. Moving Average Model (MA)

The moving average model is concerned with the accumulation of the error term in the autoregressive model, and the equation for the qth order autoregressive process defines.

$$y_t = \mu + \varepsilon_t + \sum_{k=1} \theta_i \varepsilon_{t-i} \quad (2)$$

The model is effective in eliminating random fluctuations in forecasting. The model focuses on solving for the appropriate parameter θi value with a more appropriate combination of error terms. The q in this model is similar to the p in the AR model indicating the association with the previous q time points.

3.3. Autoregressive Moving Average Model (ARMA) and Differential Autoregressive Moving Average Model (ARIMA(p,d,q))

The combination of autoregressive and moving average is defined by the formula below:

$$y_t = \mu \sum_{i=1}^{p} r_i y_{t-i} + \varepsilon_t + \sum_{k=1}^{q} \theta_i \varepsilon_{t-i} \quad (3)$$

Prediction is achieved through autoregression and moving average eliminates the parameters. In this process, the main task is to determine the two parameters p, q.

The differential autoregressive moving average model (ARIMA (p, d, q)) is an integration of the above models. AR is the autoregression, p is the autoregressive term; MA is the moving average, q is the number of moving average terms, and d is the number of differences made when the time series becomes smooth. The principle of the model is to transform a non-stationary time series into a stationary time series then the dependent variable. The model is built by regressing only its lagged values and the present and lagged values of the random error term.

4. Parameter Selection

4.1. Selection based on autocorrelation and partial autocorrelation plots

Autocorrelation coefficient (ACF) is used to determine the p-value. An ordered sequence of random variables is compared with itself, and the autocorrelation function reflects the correlation between the values of the same sequence taken at different time series with the following formula:

$$ACF(k) = \rho k\, Cov(yt, yt-k)\, Var(yt) \quad (4)$$

ρk takes values in the range of [-1, 1], with 1 denoting positive correlation and -1 denoting negative correlation. Figure 8 shows the autocorrelation property graph of a group of test data, the horizontal axis indicates the order, the vertical axis indicates the autocorrelation coefficient, and the red background on both sides of the x-axis indicates the confidence interval, which is taken as 95% here, and it is desirable if all the points after a certain order fall in the confidence interval.

Fig. 3 shows the relationship between the autocorrelation function and the order, the red background on both sides of the x-axis is the confidence interval.

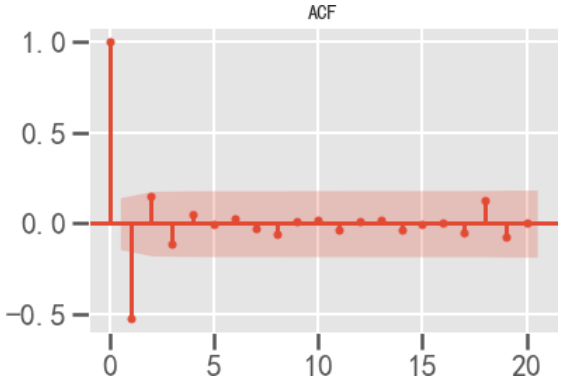

Figure 3. Plot of ACF vs. Order

The partial autocorrelation function (PACF) is used to determine the q value. For a smooth ARp model, the lag k autocorrelation coefficient pk is not a simple correlation between xt and xt-k, but xt is also affected by the k-1 random variables xt-1, xt-2, xt-k+1, which are all correlated with xt-k. Therefore, the autocorrelation coefficient pk is actually mixed with other variables on xt and xt-k. Therefore, the autocorrelation coefficient pk is actually mixed with the effects of other variables on xt and xt-k. Unlike the ACF, the PACF eliminates the interference of the middle k-1 random variables and is a strict correlation of the influence of xt-k on xt.

Fig. 4 shows the partial autocorrelation function plotted against order, with the red background on both sides of the x-axis showing the confidence intervals.

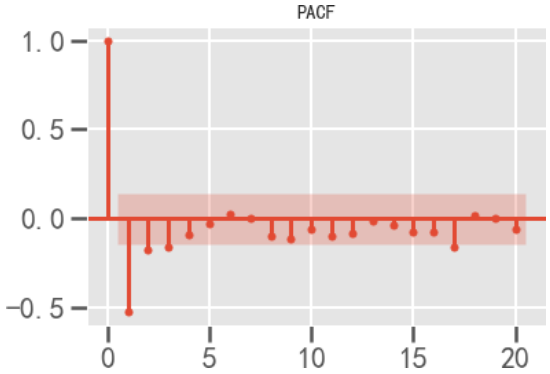

Figure 4. Plot of PAFC versus the number of orders

Referring to the selection rule of order in Table 1, we are able to determine the value of p as 4 and the value of q as 1.

Table 1. Chart for determining the order of p and q

model	ACF	PACF
AR (p)	Decay tends to zero (geometric or oscillatory)	p post-degree truncation
MA (q)	q post-degree truncation	Decay to zero (geometric or oscillatory)
ARMA (p, q)	q post-degree decay to zero (geometric or oscillatory)	p post-degree decay tends to zero (geometric or oscillatory)

But just by the above method, we may be able to determine many sets of values, and it is difficult for us to choose the most suitable one among them. At this point, another index is needed to measure the advantages and disadvantages of these selected values.

4.2. Selection based on Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC)

AIC and BIC are kind of criteria for evaluating the model, proposed by Tichi-Scale and Bayesian respectively. From the formula, it can be seen that the index is a kind of trade-off between the parameter and the accuracy of the final result, that is to say, the smaller the index is, the better it is. the smaller the value of k is, the larger the value of L is, and the smaller the entire value is.

$$AIC = 2k - 2\ln(L) \tag{5}$$

$$BIC = kln(n) - 2\ln(L) \tag{6}$$

Where k is the number of model parameters, n is the number of samples, and L is the likelihood function. p, q are larger.

The larger the value of p, q, the larger the parameter k. The difference between AIC and BIC is that BIC adds the concept of sample size, which is suitable for larger training sets.

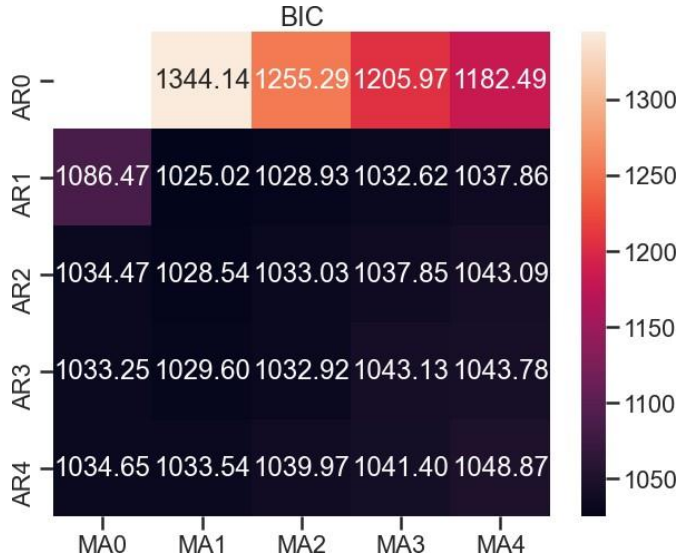

Figure 5. BIC heat map

A heat map is drawn by selecting a desirable range of AR and MA parameters (Fig. 19), and the AR and MA coordinates with the smallest values are the best p, q values.

The optimal p, q values are calculated based on AIC and BIC, respectively, where the results using the AIC criterion (p=3, q=1) are very similar to the results obtained from the autocorrelation order correlation diagram and the partial autocorrelation order correlation diagram described above (p=4, q=1).

5. Prediction results and residual test

Fig. 6 shows the prediction results of the model whose training set is the average speed per minute from 13:50-17:00 on a particular day, corresponding to the blue line in the figure. The model predicts the average speed per minute for the following 20 minutes, which corresponds to the red line in the figure. As the figure shows the comparison between the prediction results and the test set, we can find that the prediction results are more accurate in the last few minutes and deviate as the time gets longer. In this paper, we summarize the following possible reasons: (i) the amount of data in the training set

is not large enough, with less than 200 data items; (ii) the period of the values is too short, and the fluctuation of the speed per minute is too large, which will seriously affect the stability of the data.

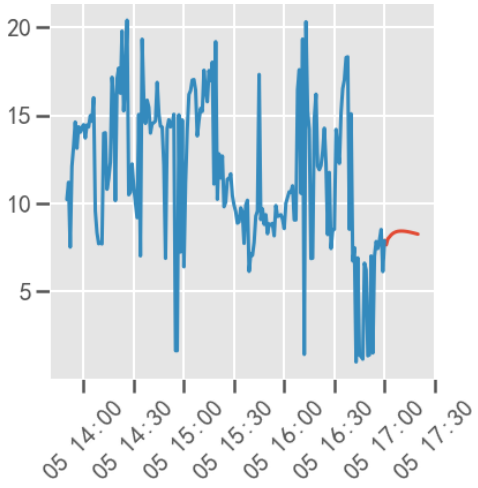

Figure 6. Prediction results

Fig. 7 consists of 4 sub-figures, sub-figure [0,0] is the intuitive distribution of the residuals line graph, the residuals can be seen uniformly distributed on both sides of the x-axis; sub-figure [0,1] represents the histogram of the residuals, as well as the normal distribution and the 01 distribution curve, which can be seen that the residuals are roughly normally distributed; sub-figure [1,0] is the QQ graph of the residuals, which is a linear distribution of the scatter, which means that the residuals are normally distributed; sub-figure [1, 1] represents the correlation coefficient of the residuals, from the table, it is known that the correlation coefficient of residuals are all within the confidence interval. 1] shows the correlation coefficients of the residuals, and the table shows that the correlation coefficients of the residuals fall within the confidence intervals. These tables show that the residuals are normally distributed, which proves that the residuals are normally distributed and are a better fit for the random errors.

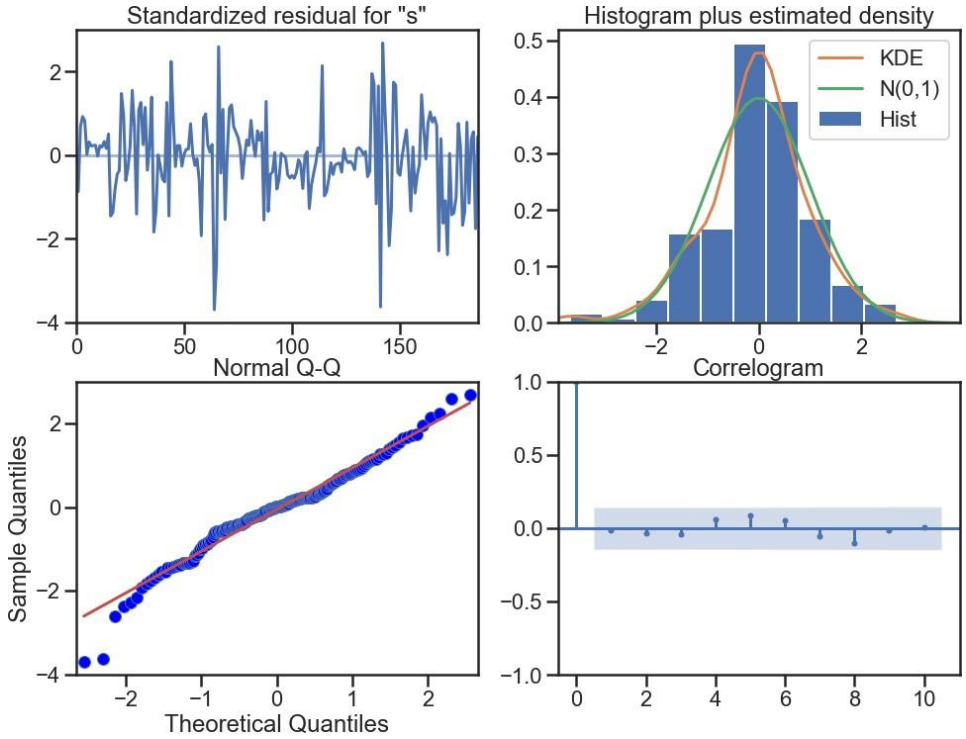

Figure 7. Residual analysis plot

6. Conclusion

Through this study, we have verified the effectiveness and practicality of ARIMA algorithm in road condition prediction, which provides new ideas and methods for road traffic management and planning. With the continuous promotion of intelligent transportation and sustainable urban development, we believe that the road condition prediction method based on the ARIMA algorithm will play an increasingly important role in the future, and make a greater contribution to the improvement of the urban transportation environment, the enhancement of transportation efficiency and the promotion of sustainable urban development. We look forward to more researchers and practitioners joining this field and working together to promote the development of intelligent transportation technology and contribute to building a more convenient, efficient and livable urban transportation environment.

In our future research, we will continue to explore road traffic prediction methods in depth, continuously optimize and improve models, explore more traffic management application scenarios, and provide more scientific and intelligent decision support for urban traffic management and planning. At the same time, we will also pay close attention to the development of intelligent transportation technology, and actively apply new technologies and methods to contribute our efforts to the development of urban transportation and sustainable planning.

References

[1] B. Chen, J. Liu, Z. Ruan, M. Yue, H. Long, and W. Yao, "Freight traffic of civil aviation volume forecast based on hybrid ARIMA-LR model," in International Conference on Smart Transportation and City Engineering (STCE 2022), M. Mikusova, Ed., Chongqing, China: SPIE, Dec. 2022, p. 69.

[2] R. Rezaiy and A. Shabri, "Drought forecasting using W-ARIMA model with standardized precipitation index," Journal of Water and Climate Change, vol. 14, no. 9, pp. 3345 – 3367, 2023.

[3] Y.-C. Jin, Q. Cao, K.-N. Wang, Y. Zhou, Y.-P. Cao, and X.-Y. Wang, "Prediction of COVID-19 Data Using Improved ARIMA-LSTM Hybrid Forecast Models," IEEE Access, vol. 11, pp. 67956 – 67967, 2023.

[4] Y. Hu, T. Peng, K. Guo, Y. Sun, J. Gao, and B. Yin, "Graph transformer based dynamic multiple graph convolution networks for traffic flow forecasting," IET Intelligent Transport Systems, vol. 17, no. 9, pp. 1835 – 1845, 2023.

[5] Z. Xia, Y. Zhang, J. Yang, and L. Xie, "Dynamic spatial–temporal graph convolutional recurrent networks for traffic flow forecasting," Expert Systems with Applications, vol. 240, 2024.