Traffic Flow Prediction Based on ARIMA Modeling

Yizhuo Lei*

Queen Mary School Hainan, Beijing University of Posts and Telecommunications, Lingshui 572423, China

*Corresponding author: jp203213787@qmul.ac.uk

Abstract. With the rapid development of urbanization and technology, the number of cars has significantly increased, resulting in increasingly serious urban problems such as traffic congestion, air pollution and frequent accidents. In order to solve these thorny problems, accurate short-term traffic flow prediction is particularly important. This study uses ARIMA model to explore the real-time, stochastic and dynamic characteristics of traffic flow. The data source of this paper is based on the detailed data set provided by spassau. The data set collects one piece of information every 15 minutes for two days, and a total of 200 records are collected. D=0 is obtained by the augmented Dickey Fuller (ADF) test, and it is inferred that the data is stable. Then ARMA (p, q) model is used for fitting and prediction. The test results show that the model performs well in the prediction of 12 steps (3 hours) in advance, the root mean square error (RMSE) is 7.00, and the average absolute percentage error (MAPE) is 11.17%. This method provides a solid technical guarantee for improving the prediction accuracy of urban traffic flow, greatly improving the management level of the city and the living standard of residents.

Keywords: Traffic Flow Prediction; ARIMA Model; Time Series Analysis.

1. Introduction

Traffic flow is a key component of urban traffic management. It takes the number of vehicles running on the road as an indicator and can be predicted according to various factors [1]. Since the initiation of reform and opening-up policies, China's economy has experienced rapid growth, accompanied by a significant rise in per capita disposable income. Consequently, automobiles have transformed from luxury items into household necessities. Concurrently, amid China's extensive infrastructure construction boom, urbanization rates have steadily increased, driving substantial migration to urban centers and fueling growing demand for private vehicles. Furthermore, the maturation of the automobile industry chain has provided consumers with greater diversity of choice and enhanced affordability, culminating in a substantial increase in automobile ownership.

According to data from the National Bureau of Statistics, by the end of 2024, China's civil motor vehicle ownership reached 352.68 million, including 6.94 million three-wheeled vehicles and slow-moving trucks. It rose by 16.51 million since last year. Of this total, the number of private vehicles reached 309.89 million, a growth of 15.62 million. Civil car ownership reached 193.43 million, an increase of 6.75 million year on year, with private car ownership totaling 182.04 million, rising by 6.64 million [2]. However, these changes have caused traffic volume to far exceed the original carrying capacity of roads, making traffic congestion an increasingly prominent issue. Traffic congestion not only reduces travel efficiency, increases travel time and costs, and weakens urban vitality, but also hinders the development of urban economy, exacerbates exhaust emissions and air pollution, and has a negative impact on social harmony [3]. With the continuous increase in the number of vehicles on the roads, efficient traffic management has become ever more pressing [4].

Accurately forecasting traffic volume has thus emerged as a pivotal prerequisite and core challenge in optimizing transportation resource allocation, alleviating congestion, and refining travel strategies [5]. Traffic flow prediction is an important component of intelligent transportation systems and has been widely applied in intelligent transportation subsystems, such as Advanced Traveler Information Systems (ATIS) and Advanced Traffic Management Systems (ATMS) [6]. Therefore, traffic flow prediction exerts multifaceted impacts on social development. First, it helps to use resources in a better way. Accurate predictions let us work with traffic departments more effectively. This allows for better planning of traffic lights and signs. As a result, the whole road network works more efficiently. Second, people in the city can plan their trips better using public predictions. If residents know the latest road conditions, they can choose to avoid the peak period and travel when the road is not so busy. This helps reduce traffic jams caused by too many vehicles for a short time.

However, traffic flow forecasting still faces many difficulties and realistic difficulties. The change of traffic flow is real-time, random and has strong instability, so it is difficult to predict. When the prediction period gets shorter, the traffic volume becomes more random and harder to predict. This makes accurate short-term traffic flow forecasting very difficult [7]. Also, prediction accuracy can be easily affected by unexpected events. For example, extreme weather can cause short-term traffic jams. Sudden accidents can block roads. Large events can lead to a sharp rise in vehicles. These unpredictable factors break the normal repeating patterns of traffic. They can even cause problems to spread to nearby roads. This makes predicting traffic even harder.

Traffic flow prediction is still a popular research topic, and many methods have been created. To understand time-related patterns, models like ARIMA, Kalman Filter, Bayesian Network, K-NN, and FNN are often used. To find location-based patterns, methods such as SAE, ITRCN, and CNN are commonly applied [8]. This study used a public traffic flow dataset from Kaggle. Data analysis was done using SPASSAU, and a time series model based on ARIMA was used to predict traffic flow accurately. The goal is to help assign traffic resources better and reduce traffic jams.

2. Method

2.1. Data Sources

The data set used in this paper was publicly available on the SPASSAU website. It was made specially for predicting city traffic flow. It includes time, location, and type details that are important for studying traffic patterns. The dataset has 200 rows and 7 columns. It covers two days, with data recorded every 15 minutes. This provides detailed information for short-term traffic flow prediction.

The dataset has several important details. The Timestamp field records the exact date and time for each 15-minute period. This helps to study how traffic changes over time. The Location field shows where the traffic sensors are placed. This makes it possible to see how traffic patterns differ from place to place. Vehicle_Count indicates the number of vehicles detected by each sensor within a given time interval. Vehicle_Speed, measured in kilometers per hour, represents the average speed of vehicles, reflecting real-time traffic conditions. Congestion_Level is an ordinal variable ranging from 0 to 5, where 0 denotes no congestion and 5 indicates severe congestion. Peak_Off_Peak, a categorical variable, distinguishes between peak and off-peak hours to support context-aware analysis. Lastly, Target_Vehicle_Count serves as the target variable for predictive modeling, representing the vehicle count in subsequent time intervals. This article focuses on analyzing urban traffic flow, retaining only Timestamp and Vehicle_Count, and the time interval is fifteen minutes. Thus, Table 1 demonstrates the replacement of Timestamp with ID, where adjacent IDs correspond to fifteen-minute intervals, and Vehicle_Count represents traffic flow status.

Table 1. Data records dictionary of urban traffic flow

Field Name	Description	Example
Id	Records the exact date and time in 15-minute intervals	1
Vehicle_Count	Number of vehicles recorded by sensors during each interval	63

2.2. Preliminary Analysis of Data

After cleaning and preprocessing the urban traffic flow data, the next step is to extract and analyze the operational characteristics of traffic flow in both temporal and spatial dimensions. Figure 1 presents the daily fluctuation graph of traffic flow corresponding to the ID during the observation period. This graph illustrates the fluctuations in traffic flow over time, revealing an overall periodic fluctuation pattern. The changes in flow for adjacent IDs (representing 15-minute intervals) reflect the short-term dynamic characteristics of the traffic. Despite these short-term variations, the overall fluctuation trend demonstrates a certain level of stability, providing a solid foundation for the subsequent fitting and prediction using the ARIMA model.

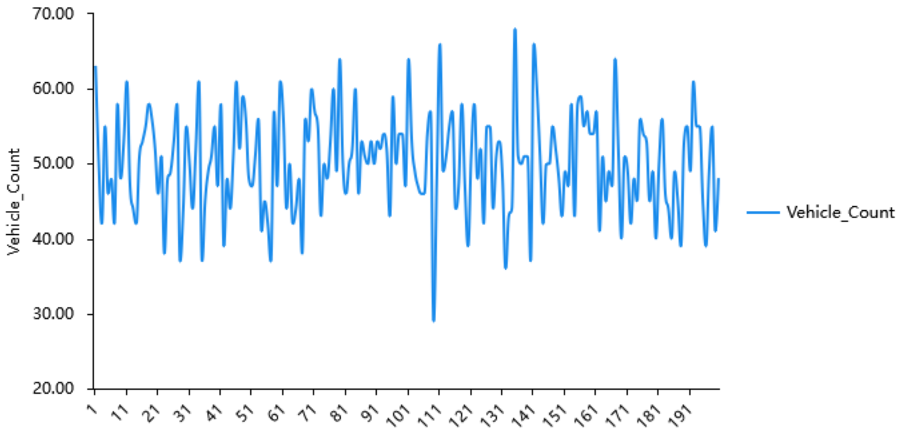

Fig. 1 data of Vehicle_Count

2.3. ARIMA Model Application

ARIMA is a very useful tool. It works well for time series data that has trends and seasonal patterns, so it is good at predicting future values. This model can play a role in many fields, for example, it can help to analyze the factors that affect consumers' decision to buy prefabricated meals [9]. In this study, based on the historical traffic flow data in the data set, ARIMA model is used for time series analysis, which is conducive to judging the potential time dependence and fluctuation law of traffic flow, and then accurately predicting the future traffic flow.

The model consists of three core components: autoregressive (AR), differential (I) and moving average (MA). In the traffic flow prediction, the autoregressive part judges the current traffic conditions through the previous traffic data, the differential part performs differential operation to make the data tend to be stable, and the moving average part is responsible for improving the prediction accuracy. This combination enables ARIMA model to effectively capture the change characteristics of traffic data and achieve accurate prediction of future traffic conditions.

ARIMA model contains three parameters: p, d and q, where p represents autoregressive order, d represents differential order, and q represents moving average order. Each of these three parameters has its own responsibility. Reasonably determining the values of these three parameters is the key to building an effective ARIMA model.

3. Results and Discussion

3.1. Data Stationarity Verification

The stability of time series data is a necessary prerequisite for using ARIMA model, which means that the average value and distribution of data are relatively concentrated and will not change significantly with time. Augmented Dickey Fuller (ADF) program evaluates the stationarity of variables by detecting the unit root [10]. Therefore, this study uses ADF test to judge the stationarity of traffic flow data, and carries out visual auxiliary analysis according to the scatter diagram.

3.1.1 ADF Test Results

The Augmented Dickey-Fuller (ADF) test was conducted with a difference order of 0. The test statistic was -14.049, which is considerably smaller than the critical value of -3.464 at the 1% confidence level, and the p-value was 0.000 ($p < 0.05$). This significant result allows us to reject the null hypothesis of the existence of a unit root (non-stationarity), confirming that the Vehicle_Count series is stationary. As a result, no differencing is required (i.e., the difference order is d = 0), and the ARIMA(p, 0, q) model can be directly applied, simplifying the process of model parameter selection.

The results of the ADF test are summarized in Table 2 below. Additionally, the scatter plot of the data further supports the stationarity of the series. The values of Vehicle_Count are concentrated within the range of 35-65 and are evenly distributed without showing any significant upward or downward trends. The data points exhibit stable vertical fluctuations with no signs of "funnel-shaped" heteroscedasticity, indicating constant variance, which aligns with the intuitive expectations for a stationary series.

Table 2. Result of Vehicle_Count - ADF Test

Order of difference	**t-statistic**	**p-value**	**Critical Value (1%)**	**Critical Value (5%)**	**Critical Value (10%)**
0	-14.049	0.000	-3.464	-2.876	-2.575

3.1.2 Scatter Plot Evidence

The scatter plot of Id and Vehicle_Count shown in Fig. 2 further confirms the stationarity of the sequence from a visual perspective: the traffic flow values are concentrated in the range of 35-65, and are evenly distributed overall, with no obvious upward or downward trend; the vertical fluctuation range of the data points remains stable, and there is no "funnel-shaped" heteroscedasticity pattern that significantly expands or shrinks with the increase of Id, indicating that the variance of the sequence is constant, which meets the intuitive judgment criteria for stationarity.

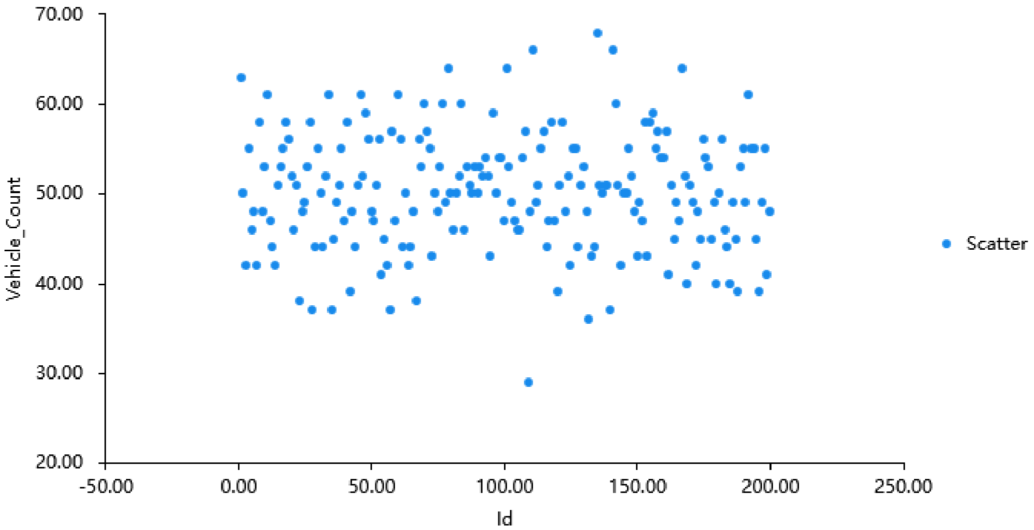

Fig. 2 Scatter plot of data of Vehicle_Count

3.2. ARIMA Model Construction and Adaptability Analysis

Based on the stationarity of the verified sequence (d=0), the model construction can be simplified to the ARMA(p, q) form (this is a special case of the ARIMA model). Its adaptability is mainly reflected in two aspects: time-dependent modeling and the simplicity and robustness of the model itself. In terms of time-dependent modeling, the inherent temporal autocorrelation of traffic flow (such as the high correlation of traffic flow in adjacent time periods), combined with the rapid decay characteristics of the stationary sequence ACF and PACF shown as Fig.3 and Fig.4, provides a basis for scientific identification of the AR order (p) and MA order (q). High-frequency sampling data (such as shorter time intervals) further strengthens the correlation between the lag terms in adjacent time periods, laying a solid foundation for AR term modeling. In terms of simplicity and robustness, low-order ARMA models (such as p=1, q=1) are usually sufficient to capture the main fluctuation patterns of the sequence. This not only significantly reduces the complexity of parameter estimation, but also effectively avoids the risk of overfitting and improves the generalization ability of the model.

Automatic model identification in SPSSAU, utilizing the ACF and PACF plots for Vehicle_Count (Fig. 3 and Fig. 4), yielded optimal ARIMA parameters of p=6 (autoregressive order) and q=7 (moving average order).

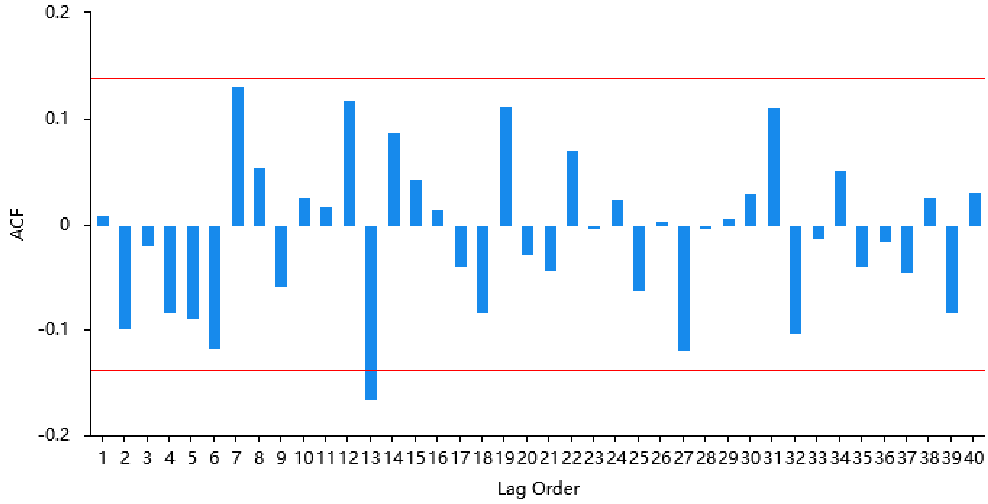

Fig. 3 plot of ACF

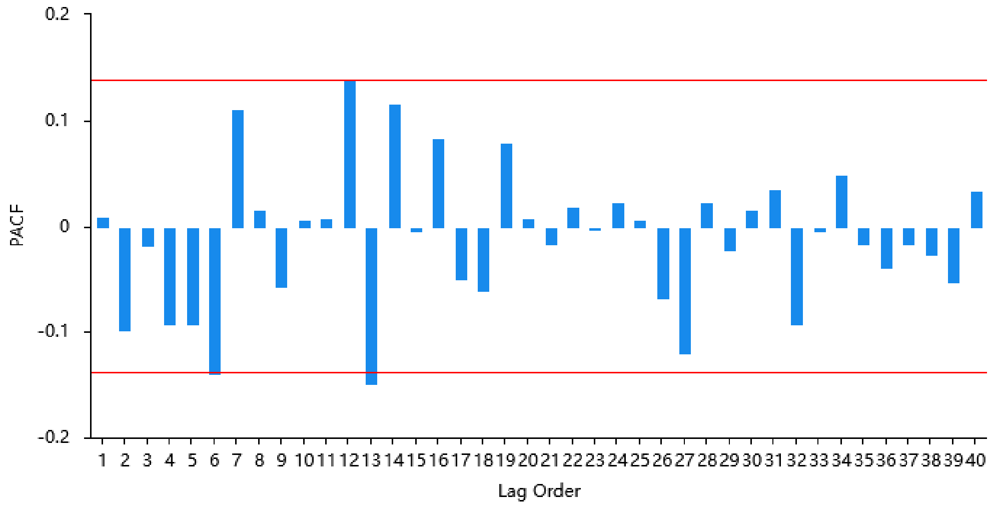

Fig. 4 plot of PACF

3.3. Prediction Results and Model Performance Evaluation

The ARIMA model was used to forecast traffic flow for the next 12 time periods, with the model's performance assessed using several error metrics: RMSE = 6.2139, MSE = 38.6127, MAE =4.9034, and MAPE = 10.19%. These results indicate a satisfactory level of prediction accuracy. Specifically, the RMSE value of 6.2139 suggests that the average deviation between the predicted and actual observed values is approximately 6 vehicles. The MAPE value of 10.19% indicates a relatively small relative error, further confirming the model's high accuracy.

As shown in Table 3, the predicted traffic flow for the next 12 time periods is presented, facilitating a detailed comparison of the forecasted and actual values. The data in table 3 further support the ARIMA model's accuracy in predicting short-term traffic flow changes.

Table 3. Predicted value (12 periods)

Fore cast	Back ward 1	Back ward 2	Back ward 3	Back ward 4	Back ward 5	Back ward 6	Back ward 7	Back ward 8	Back ward 9	Back ward 10	Back ward 11	Back ward 12
Valu e	54.4 62	50.3 08	50.3 97	51.3 12	52.4 88	49.9 45	47.6 32	51.1 99	50.5 14	49.8 13	48.4 38	50.4 28

Fig. 5 demonstrates the ARIMA(6,0,7) model's strong performance through close alignment between fitted/predicted values and historical Vehicle_Count trends, with accurate tracking of short-term fluctuations confirming its effectiveness in capturing traffic dynamics. The narrow 95% confidence interval surrounding predictions indicates precise uncertainty quantification, while quantitative validation is evidenced by exceptional forecasting accuracy: a Mean Absolute Percentage Error (MAPE) of 10.19%, complemented by Root Mean Squared Error (RMSE = 6.2139), Mean Absolute Error (MAE = 4.9034), and Mean Squared Error (MSE = 38.6127). This sub-11% MAPE reflects high traffic volume prediction reliability, with consistent performance across all error metrics - combined with robust parameter estimation (p=6, q=7) and adherence to intrinsic data patterns - collectively validating the model's operational credibility and robustness for traffic flow forecasting.

In summary, these analyses confirm that the ARIMA model is effective in capturing traffic flow fluctuations and providing accurate short-term forecasts. The insights derived from this model will support urban traffic management, particularly in optimizing traffic scheduling and planning, thus showcasing its practical application in real-world scenarios.

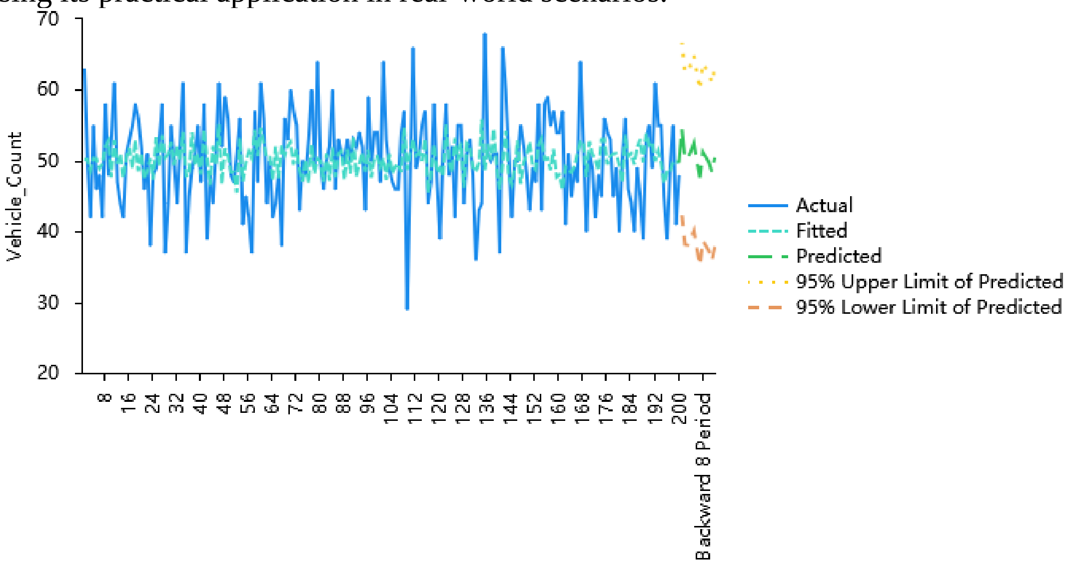

Fig. 5 plot of vehical_count model fitting and prediction

4. Conclusion

Based on the number of cars in the urban traffic flow data set published on the kaggle platform, this paper analyzes the data and makes a preliminary short-term traffic flow prediction. The research method uses the very classic ARIMA model, and determining the three parameters of P, D and Q is the key link of using the model. Therefore, the ADF test is carried out on the data set, and the data of the data set is stable. At the same time, the visual scatter diagram is used for auxiliary analysis, which shows that the data is evenly distributed and corresponds to the results of ADF test, so d=0 is obtained. Then the characteristics of stationary series are observed, and the values of P and Q are determined in combination with the requirements of autoregressive model (AR) and moving average model (MA); Finally, the most suitable ARIMA model is determined by parameter setting.

The results of this study are of great significance to urban traffic management. By accurately predicting urban traffic flow, the transportation department can better manage road traffic, help optimize resource allocation, ease traffic congestion, shorten travel time, so as to provide more convenient travel experience for urban residents.

In summary, ARIMA model can provide valuable data support for urban planning and traffic management, and can effectively identify the time regularity in traffic data, so it is very suitable for short-term prediction. In future research, more factors that may affect traffic flow need to be considered, such as weather, large-scale activities, etc. At the same time, the model can also be combined with more advanced machine learning to obtain more accurate prediction and wider application range.

References

[1] Prasanna S. et al. Traffic-Flow Predictor: Predicting Traffic Volume and Patterns. 2025 International Conference on Intelligent Computing and Control Systems (ICICCS), Intelligent Computing and Control Systems (ICICCS), 2025 International Conference on, [s. l.], p. 1151–1155, 2025.

[2] National Bureau of Statistics. Statistical Communiqué on National Economic and Social Development of the People's Republic of China in 2024. China Information News, 2025-02-28 (001).

[3] Xiao L L. Research on Traffic Volume Prediction Method Based on Improved Autoformer. Dalian Jiaotong University, 2025.

[4] Yang X. et al. Parallel Self-Learned and Predefined Joint Spatial–Temporal Graph Convolutional Networks for Traffic Flow Prediction. IEEE Internet of Things Journal, Internet of Things Journal, IEEE, IEEE Internet Things J, [s. l.], v. 12, n. 9, p. 11698–11707, 2025.

[5] Lou W H. Research on Short-term Traffic Flow Prediction Method Based on Multivariate LSTM. Transportation Enterprise Management, 2025, 40(04): 97-99.

[6] Liu M Y, Wu J P, Wang Y B, He L. Traffic Flow Prediction Based on Deep Learning. Journal of System Simulation, 2018, 30(11): 4100-410.

[7] Rui L L, Li Q M. Short-term traffic flow prediction algorithm based on combined model. Journal of Electronics & Information Technology, 2016, 38(5): 1227-1233.

[8] Liu C. Research on Traffic Flow Prediction Based on Deep Learning. Tianjin Polytechnic University, 2023.

[9] Zhang Z. et al. Global Temperature Analysis Based on ARIMA Model and Pearson's Correlation Coefficient. 2025 International Conference on Electrical Drives, Power Electronics & Engineering (EDPEE), Electrical Drives, Power Electronics & Engineering (EDPEE), 2025 International Conference on, EDPEE, [s. l.], p. 443–448, 2025.

[10] Zhang F. Zhao Y. Zhang S. Wu W. & Tan C. Spacecraft Equipment Health Condition Monitoring Based on Augmented Dickey-Fuller Test and Gaussian Mixture Model. 2021 IEEE International Conference on Mechatronics and Automation (ICMA), Mechatronics and Automation (ICMA), 2021.