Research On the Forecasting of Sales Volume of New Energy Vehicles Based on SARIMA Model

Jiageng Chen #, Yujun Huang #, Han Jin *, Xiaofeng Chen

School of Civil Engineering and Transportation, Guangzhou University, GuangZhou, China, 510006

* Corresponding Author Email: 13352870289@163.com

#These authors contributed equally.

Abstract. In recent years, the new energy vehicle market has experienced rapid growth due to policy support and environmental enhancements; however, sales volume fluctuates due to various factors, making precise forecasting essential for production planning, supply chain management, and market strategy. This paper conducts a comprehensive forecasting analysis of new energy vehicle sales volume utilizing the SARIMA model, including the fluctuations in the worldwide new energy vehicle industry. The research gathers and preprocesses monthly sales volume data of new energy cars from January 2022 to April 2024 and develops a SARIMA prediction model by thoroughly considering the trend, seasonal, abrupt, and stochastic error components. The projections indicate that the sales volume of new energy cars will attain 478,000 units in May 2024, increase to 883,000 units in June, and see a minor decline to 588,000 units in July, while still being elevated. This particular numerical conclusion offers a scientific foundation for the future advancement of the new energy vehicle market and is highly significant in directing the optimal placement of charging facilities, assessing the adequacy of existing charging infrastructure, and recommending optimal locations for charging station installations.

Keywords: SARIMA model, new energy vehicle sales, Time series forecasting, Trend and Seasonality Analysis, Forecasting Accuracy.

1. Introduction

Recently, global demand for new energy vehicles has surged, particularly in the Chinese market, where sales have consistently increased annually. The growth trajectory of the new energy vehicle market is not constant, and its sales volume exhibits volatility influenced by various factors, including policy alterations, seasonal variations, market supply and demand dynamics, and fluctuations in the external economic landscape [1-2]. Consequently, correct forecasting of sales volume for new energy vehicles is crucial for developing effective manufacturing plans, enhancing supply chain management, and executing targeted market strategies. Conventional time series models, such as ARIMA and moving average models, while capable of capturing trends and volatility to some degree, frequently overlook seasonal variations in the data [3-5].

The study initially gathers and preprocesses new energy sales data sourced from public market reports, documenting the monthly sales of new energy vehicles from January 2022 to April 2024. Data preprocessing encompasses three components: data cleansing, format normalization, and trend identification to guarantee the precision and relevance of the data. The constraints of traditional algorithms for time series forecasting were further examined, followed by an analysis of trends and seasonality, leading to the selection of the Seasonal Autoregressive Integrated Moving Average (SARIMA) model for forecasting. The parameters of the SARIMA model were initially selected by result analysis, and the seasonality parameters were established. The algorithm was ultimately calibrated utilizing previous data and readied to project revenues for the forthcoming three months.

2. Data sources and pre-processing

2.1. Data sources

This study utilizes data on new energy vehicle sales obtained from www.gotohui.com compiled by certified market research firms to guarantee the accuracy and timeliness of the information. The dataset documents monthly sales of new energy vehicles from January 2022 to April 2024, encompassing 28 data points, each indicating total national sales for one month. The data have an extensive temporal range and exhibit variations throughout numerous seasons, rendering them highly appropriate for time series analysis.

2.2. Data pre-processing

Data preprocessing is an important part of data analysis that directly affects how well the next model is built and how accurate the predictions are. The main tasks that go into preparation are cleaning the data, making sure the format is correct, and finding trends.

2.2.1 Data Preprocessing

In the data preprocessing phase, the initial step involved data cleansing, where a thorough review confirmed the absence of missing values or outliers in the dataset. Following this, the format standardization process was implemented, during which dates originally in text format (e.g., "January 2022") were converted into a Python-recognizable date format (e.g., 2022-01-01). This transformation was essential to enable the use of time series analysis tools, thereby allowing the study to more effectively identify and analyze trends and seasonal variations within the data.

2.2.2 Trend detection

The time series of new energy vehicle sales, depicted in Figure 1, demonstrates a notable upward trend in sales. Despite certain fluctuations, sales numbers generally increase with time. The graph displays numerous substantial changes, particularly in early 2023 and early 2024. These variations may relate to market seasonality, promotional activities, or external economic events.

Annual fluctuations in sales quantities indicate that sales may be affected by seasonal variables. Sales seemingly reach their zenith in the fourth quarter, potentially attributable to year-end automotive purchase discounts and holiday marketing. A year-over-year analysis reveals that specific months in 2023 and 2024 exhibit increased sales volumes relative to 2022, signifying overall market expansion.

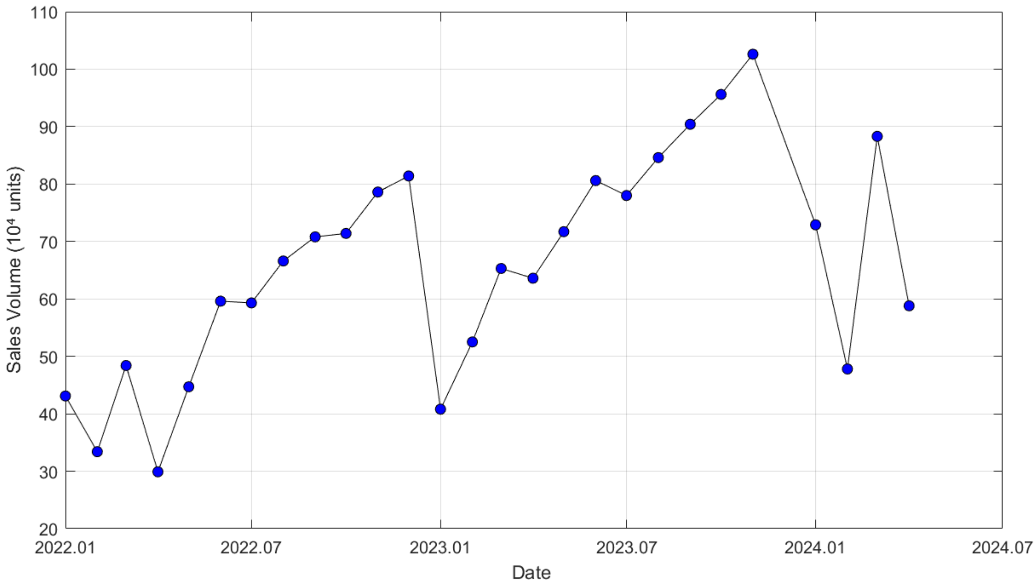

Figure 1. Time series diagram of new energy vehicle sales

2.3. Statistical tests for time series analysis

2.3.1 Smoothness test

Smoothness is an essential need for time series analysis. Non-smooth series may lead the model to misinterpret the data's characteristics, hence impacting the accuracy of forecasting outcomes. The research used the Augmented Dickey-Fuller (ADF) test to assess the stationarity of the series. The ADF test is a widely employed statistical procedure to ascertain if a specific time series possesses a unit root, indicating that it is a non-stationary series.

2.3.2 Autocorrelation and partial autocorrelation analysis

The autocorrelation function (ACF) and partial autocorrelation function (PACF) are essential instruments for detecting patterns, such as trends and seasonality, in time series data and for establishing the parameters of the ARIMA model [6-7]. The ACF illustrates the correlation between a time series and its historical values, but the PACF demonstrates the direct correlation between the time series and its historical values, without the influence of intermediate values. Figure 2 displays the findings of the ACF and PACF.

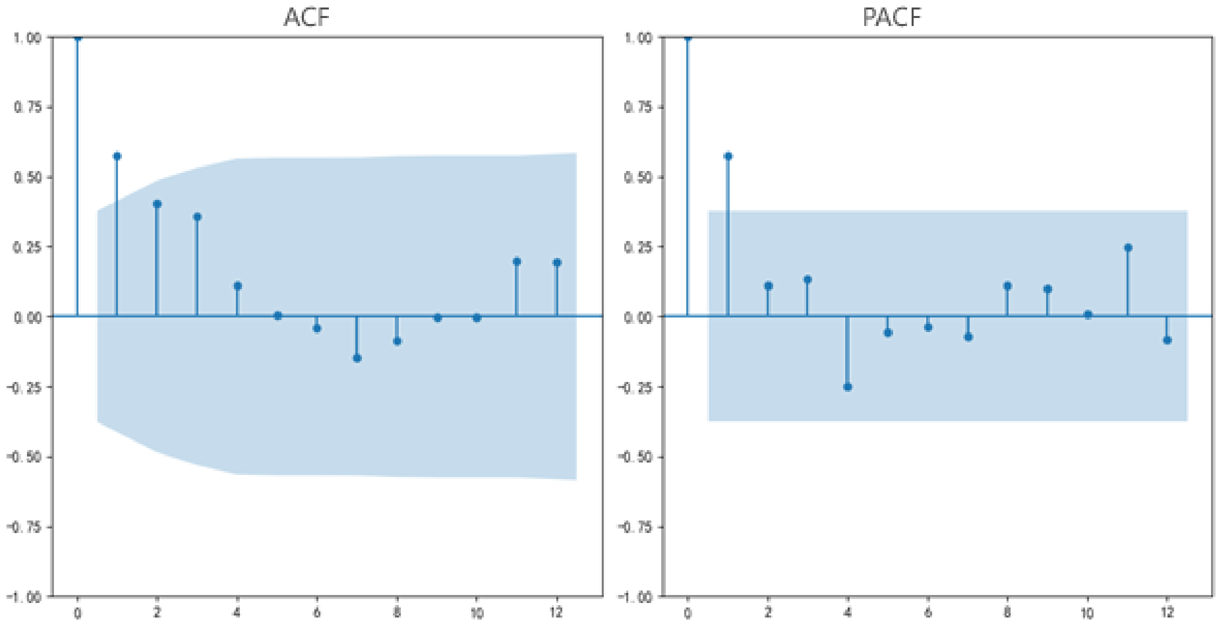

Figure 2. Autocorrelation and partial autocorrelation analysis

The ACF plot indicates that the series exhibits a prolonged autocorrelation, typically indicative of non-stationary time series and potentially suggesting seasonal influences. The importance of the first lag in the PACF indicates that a single autoregressive term may suffice to accurately represent this series, as the skewed autocorrelation coefficients for delays after the first rapidly converge to zero.

3. Modelling and parameter adjustment

3.1. Modelling

Time series forecasting methods primarily encompass trend component forecasting, seasonal component forecasting, burst component forecasting, and random error component forecasting. The Holt-Winters algorithm, as an illustration of traditional time series forecasting, incorporates smoothing parameters for the level, trend, and seasonal components. The standard algorithm is unsuitable for long-term analysis of extensive data sets because these parameters, once established, are immutable and require multiple attempts to ascertain the ideal value [8]. Only short-term predictions yield superior results; as the prediction time extends, the prediction error increases significantly. Therefore, long-term data forecasting does not meet the necessary standards and necessitates the application of a new time series prediction model for analysis. The analysis revealed evidence of non-stationarity and pronounced seasonal variations in the data. Consequent to these findings, the study selected the Seasonal Autoregressive Integrated Moving Average (SARIMA) model for forecasting purposes. The SARIMA model proficiently manages time series data exhibiting both seasonal and non-seasonal variations [9-11].

3.2. Parameter adjustment

The parameters of the SARIMA model (p, d, q) and the seasonal parameters (P, D, Q, s) must be established, and the SARIMA model parameters should be optimized through a grid search approach to identify the optimal combination that minimizes prediction error.

3.2.1 Prediction of trend components

Initially, data analysis was conducted to examine the historical trends in monthly sales volume of new energy vehicles. The historical data pertaining to initial sales volume and the slope of each segment were fitted to a linear model, ensuring continuity between each fitted line. The historical sales volume data served as a training sample for modeling, subsequently establishing the relationship between the historical sales volume of energy vehicles and the trend line, thereby deriving the predictive model for the trend component. Extract the trend components. Forecasting model.

$$T(\mathrm{t}) = f\left\{X_k, Slope_k\right\} \tag{1}$$

$$K_{T+\mathrm{i}}^{'} = Max\left\{K_{T+i}, \gamma \cdot \min\left\{K_T, K_{T-1}, \cdots, K_{T-N+1}\right\}\right\} \tag{2}$$

In the aforementioned equation, $K_{T+i}^{'}$ represents the slope of improvement post-compensation; if the latest consecutive N, the slope of a number, is not less than zero, then the slope must not be less than zero; γ is modifiable until an ideal constant is achieved.

3.2.2 Forecast of seasonal components

The periodicity of sales volumes for energy vehicles was established by a frequency analysis of historical sales data. A forecasting model for the seasonal component was developed by calculating the average bloom time for the corresponding period each year, employing methods such as seasonal differencing and least squares. The initial aspect to be determined in the study is the cycle time, while statistical analysis is employed to extract characteristics from the extensive data and execute the difference operation to derive matrix A.

$$A = \begin{bmatrix} x_2 - x_1 & x_3 - x_2 & \cdots & x_{n-1} - x_{n-2} \\ x_3 - x_1 & x_4 - x_2 & \cdots & x_n - x_{n-2} \\ \vdots & \vdots & \vdots & \vdots \\ x_m - x_1 & x_{m+1} - x_2 & \cdots & x_{n+m-3} - x_{n-2} \end{bmatrix} \tag{3}$$

The linear fitting of each row of matrix A produces distinct fitted lines represented by Y=aX+b, with the count of rows exhibiting the minimal fitting error indicating the period. L.p. signifies the quantity of samples within each period L. The seasonal component at each point q (q=1, 2, p) can be articulated as the mean of the data at the same position q over the p samples, and from this relationship, the seasonal component can be deduced.

$$S_{pi}(t) = \sum_{q=1}^{p} X_{qi} / p \tag{4}$$

3.2.3 Prediction of Burst Component B

Burst Component B typically emerges from unforeseen occurrences, pinpointing unexpected events that could influence the sales volume of new energy vehicles, including relevant regulatory alterations. Generally, the contingency components are quantifiable; specifically, the contingency component B related to the sales volume of energy vehicles can be articulated in relation to a distinct category associated with a certain value.

$$B(t)_{value} = \{Burst_{v1}, Burst_{v2}, \cdots, Burst_{vn}\} \tag{5}$$

$$B(t)_{type} = \{Burst_{t1}, Burst_{t2}, \cdots, Burst_{tn}\} \tag{6}$$

$$B(t) = \{CELLID, Burst_{v1}, Burst_{t1}, \ldots, Burst_{vn}, Burst_{tn}\} \tag{7}$$

In the study and projection of energy vehicle sales volume, it is essential to identify the relevant burst component B(t) for each energy vehicle type and incorporate it into the forecasting equation.

3.2.4 Prediction of the random error component R

Examine the residuals in the historical data, specifically the disparity between the actual observed values and the anticipated values derived from the trend and seasonal component models. In the big data prediction model, the random error component is not independently distributed; rather, its predicted value is derived by subtracting the trend, seasonal, and burst components from the historical data of energy vehicle sales volume. The treatment outcome guarantees that the random error components are more accurate.

3.2.5 Forecast of Energy Vehicle Sales

Forecasting future energy vehicle sales is done using the formula:

$$X(t) = (1 + B(t)) \times (T(t) + S(t) + R(t)) \tag{8}$$

4. Parameter setting and predictive analysis of SARIMA model

4.1. Parameter setting

Following the aforementioned analysis and forecasts, the study initially identified the parameters of the SARIMA model. In light of the data's non-stationarity, the study initially applied first-order differencing. Consequently, due to the pronounced beginning lag observed in the PACF charts, the autoregressive term (p) was selected as 1, the differencing term (d) as 1, and the moving average term (q) as 1. Furthermore, due to the data demonstrating considerable annual regularity, the study established the seasonality parameter with a duration of 12 months. The seasonal autoregressive term (p), seasonal differencing term (d), and seasonal moving average term (q) are all established at 1 to account for seasonal fluctuations in the data.

4.2. Model predictions

The study employed the SARIMA model based on the aforementioned parameters. The study employed a substantial amount of historical data as the training set while reserving the most recent months as the test set to evaluate the model's predictive capability. Upon completion of the model fitting, the study forecasted sales for the subsequent three months. The model's validity and accuracy were confirmed by juxtaposing the predictive outcomes with the actual sales data. The model's predictive outcomes indicate that the sales volume of new energy vehicles will persist in its growth, aligning with market trend expectations.

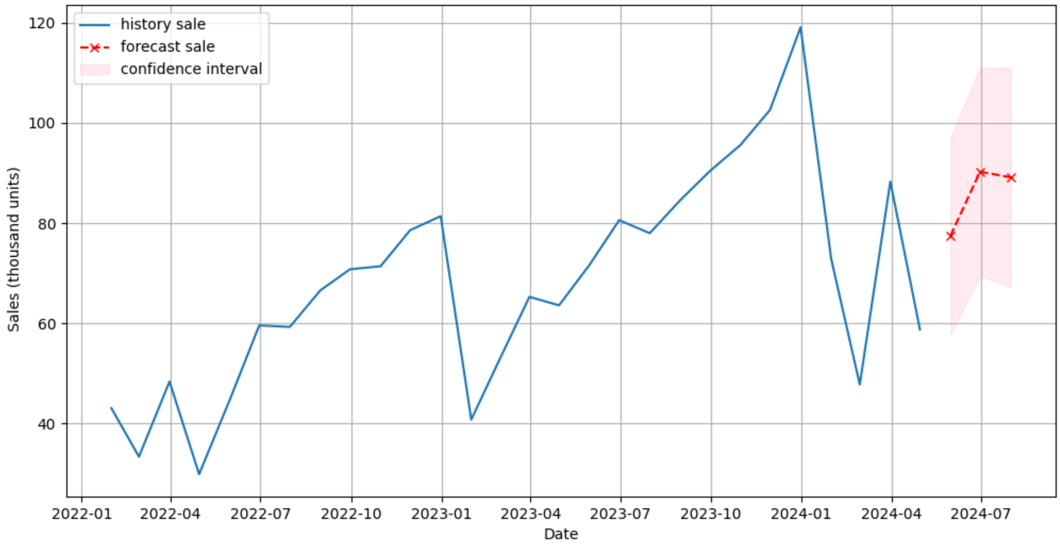

Figure 3. Sales forecast for the next three months

Figure 3 illustrates projections for new energy vehicle sales from May to July 2024, indicating an upward trend during this period. The projections include 770,000 for May, 900,000 for June, and 890,000 for July, reflecting a significant rise in sales during the summer period. The extensive confidence intervals for the projections, especially in June and July 2024, signify a degree of uncertainty in the forecasts. This may arise from the significant volatility of historical data, rendering the model more unpredictable in forecasting future patterns.

5. Conclusions

This study offers a comprehensive predictive analysis of new energy vehicle sales volume utilizing the SARIMA model. The study effectively constructed a SARIMA forecasting model by meticulously gathering and preparing monthly sales volume data for new energy cars from January 2022 to April 2024, incorporating a trend component, a seasonal component, a sudden component, and a random error component. The study conducted comprehensive projections of the trend and seasonal components during model creation, thoroughly accounting for unforeseen events and other stochastic factors that could influence sales volume. The projection indicates that the sales volume of new energy cars would persist in its growth over the next three months, aligning with market expectations. This study establishes a scientific foundation for the future advancement of the new energy vehicle market, aids in the strategic arrangement of charging infrastructure, evaluates the adequacy of current charging facility configurations, and recommends the best locations for charging station installations.

Notwithstanding the findings of this investigation, certain deficiencies persist. The study utilized sales data exclusively up to April 2024, thereby affecting the accuracy and timeliness of the estimate. To enhance the study, three aspects can be supplemented and refined in the future: first, to perpetually gather and update sales data of new energy vehicles to improve the accuracy and timeliness of forecasts; second, to investigate additional factors influencing the sales volume of new energy vehicles, such as consumer preferences and technological advancements, and integrate them into the forecasting model; and third, to incorporate advanced time series forecasting methods or machine learning algorithms to optimize the SARIMA model, thereby enhancing the accuracy and stability of the forecasts. This study illustrates the complete process from data collection, preprocessing, model selection, and parameter optimization to the final forecasting analysis, thereby validating the method's effectiveness and feasibility while offering novel insights for research in the domain of time series forecasting.

References

[1] Jiang Jiachen, Mei Yuxin, Pan Yaoyao, Wang Jie. Development Assessment and Sales Forecast of New Energy Vehicles [J] Journal of Taizhou College. 2024; 46(03):9-15.

[2] Li Wenyuan. New Energy Vehicle Sales Forecast and Influencing Factors Analysis [J] Operator. 2024; (12):92-93.

[3] Wang SK, Liu JH. Research on new energy vehicle sales forecasting based on time series analysis [J] Statistics and Management. 2022; 37(07):88-97.

[4] Zou, Rui, Liu, Jihua, Xu, Si-wei. Forecasting sales volume of new energy vehicles based on time series - A case study of BYD [J] Technology and Industry. 2024; 24(15):87-93.

[5] Yu Mingyang, Shen Bin. Analysis of New Energy Vehicle Sales under Variable International Circumstances - A Compound Forecasting Model Based on Sudden Factors [J] China Business Journal. 2023; (12):164-168.

[6] Xie Jiayin. New energy vehicle SUV sales forecast based on ARIMA promotes the formation of new quality productivity [J] China Economic and Trade Review. 2024; (10):136-138.

[7] Chen Kexiu, Liu Juan. Forecasting new energy vehicle sales of Ola Black Cat based on ARIMA model [J] Modern Industrial Economy and Informatisation. 2022; 12(03):169-171.

[8] Ouyang Di, Cui Jia. Spatial and temporal dynamics of air pollutants in Harbin under the carbon neutral strategy based on the Holt-Winters time series model [J] Environmental Protection Science. 2023; 49(02):112-119.

[9] Wang Ying, Han Baoming, Zhang Qi, Li Dewei. Prediction of inbound passenger flow of Beijing underground based on SARIMA model [J] Transportation Systems Engineering and Information. 2015; 15(06):205-211.

[10] Sun Xianghai, Liu Tanqiu. Short-term Traffic Flow Prediction on Urban Roads Based on SARIMA Model [J] Highway Traffic Science and Technology. 2008; (01):129-133.

[11] Li Jie, Peng Qiyuan, Yang Yuxiang. Passenger flow prediction of Guangzhou-Zhuhai Intercity Railway based on SARIMA model [J] Journal of Southwest Jiaotong University. 2020; 55(01):41-51.