Academic Journal of Science and Technology
ISSN: 2771-3032 | Vol. 12, No. 2, 2024

Analysis of New Energy Vehicle Development in China Based on LSTM and ARIMA Models

Yuhe Zhang *

College of Computer Science and Engineering, Dalian Minzu University, Dalian, China
*** Corresponding author Email:** 1637340371@qq.com

Abstract: This study comprehensively analyses the development factors of new energy vehicles in China based on data and literature, adopts principal component analysis to reduce the dimensionality to deal with the influence indicators, and combines multiple regression and LSTM model to predict the development trend of the industry. For the relationship between electrification and the ecological environment, multiple linear regression and ARIMA methods are tried to explore the impact of carbon emissions on the ecological environment. These methods provide a comprehensive perspective for a comprehensive understanding of the development of new energy vehicles and provide a reference basis for future industrial development and ecological environment improvement.

Keywords: Multiple Regression Modelling; LSTM; ARIMA.

1. Introduction

In this study, we explored the current status and trends of the development of China's new energy vehicle sector and predicted the development of China's new energy electric vehicle industry in the next 10 years by using a multiple regression prediction model and an LSTM model. We also used the ARIMA method to forecast carbon emissions to reflect the impact of electrification on the ecological environment. Through the comprehensive use of data analysis and literature research, we provide a comprehensive and objective analysis of the various factors affecting the development of new energy vehicles, ranging from economic factors and costs, market and penetration, infrastructure and innovation to social macro indicators, which provides an important reference for in-depth understanding and promotion of the development of the new energy vehicle industry.

2. Determination of Indicators

2.1. Selection of Indicators

Before solving the problem, we need to identify indicators as factors affecting the development of new energy electric vehicles in China. So, we found nearly 25 related papers on China National Knowledge Infrastructure, Google Scholar website and various journal websites, counted the number of times various factors appeared in the articles, and obtained the following categories through sorting: economic factors and cost categories, market and penetration category, infrastructure and innovation category, and social macro-indicator category. As shown in Figure 1 below is the indicator evaluation system we proposed.

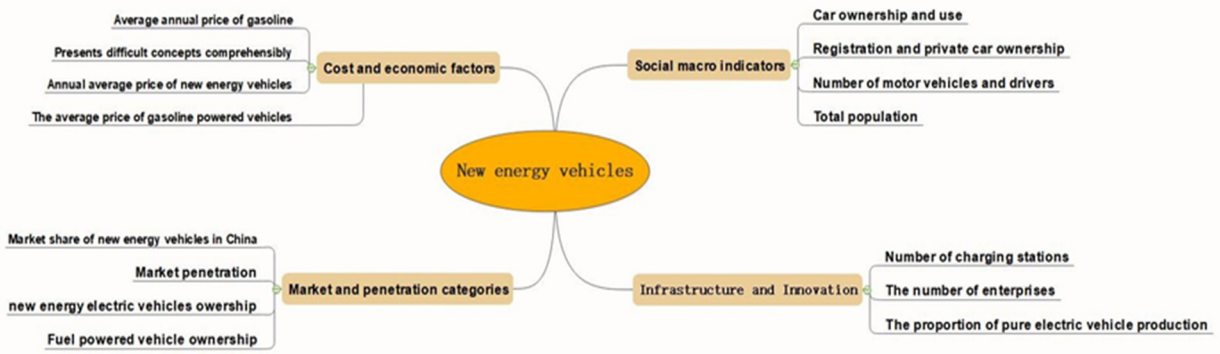

Figure 1. Indicator evaluation system

2.2. Data Preprocessing-outliers

The Kolmogorov-Smirnov test method is referred to as the K-S method. It is tested by using the maximum distance between the distribution function of the sample and the assumed theoretical distribution function [1]. Generally speaking, it is simpler than the chi-square test. When the maximum difference is less than a certain threshold, we treat this data as fitting this distribution. In the case where the gap between the two is small, it is inferred that the sample was taken from a certain distribution we know. The test statistic is:

$$Z = \sqrt{n}\max\left(|F_n(x_{i-1}) - F(x_i)|, |F_n(x_i) - F(x_i)|\right) \quad (1)$$

When H is true, the distribution of Z converges to the Kolmogorov-Smirnov distribution. That is, when the sample is taken from a one-dimensional continuous distribution:

$$F = sup|F_n(x_i) - F(x_i)| \quad (2)$$

The Kolmogorov-Smirnov test will finally get a p value. If the p value is less than the significance level, the null hypothesis needs to be rejected, that is, the two samples do

not belong to the same distribution. The Kolmogorov-Smirnov test was performed on some data to determine its distribution pattern. The results of the Kolmogorov-Smirnov test analysis are shown in Table.1. below.

Table 1. Kolmogorov-Smirnov test result

Indicator	P
Fuel price (yuan/liter)	0.74871904
Electric vehicle charging cost (yuan/kWh)	0.99619416
Average price of electric vehicles (10,000 yuan)	0.998921658
Average price of fuel vehicles (10,000 yuan)	0.999669639
Amount of government subsidies for new energy vehicles (100 million yuan)	0.997951246

We can see from the table that the calculated p values meet the requirements of normal distribution, so most of the data obeys normal distribution. Therefore, principles can be introduced to determine outliers at this time. When the given data set obeys the normal distribution, 99.7% of the data falls within three standard deviations of the mean. The formula for determining outliers is as follows:

$$P(|x-\mu|>3\sigma)\leq 0.003 \quad (3)$$

By analyzing the actual situation and artificial judgment in Table.2, we came to the conclusion: in the collected data set, the data on the market penetration rate of fuel vehicles in 2019 was an outlier, and we chose to replace it with missing values.

Table 2. Analyze result

Years	Traditional fuel vehicle market penetration rate (%)	New energy vehicle energy efficiency (km/kWh)
2019		6
2020	94.3	6.3
2021	93.6	
2022	92.7	6.9

2.3. Data Preprocessing-Dimensionality Reduction

Through principal component analysis [2], the economic factors and cost indicators are dimensionally reduced, and the results are shown in Table.3.

For infrastructure and innovation categories that failed the test, a correlation analysis model was established, and person correlation was introduced. Determine the relationship between indicators.

In order to increase the visualization of the results, MATLAB was used to draw a matrix heat map of the correlation analysis, as shown in Figure 2.

Table 3. Dimensionality reduction results

Years	New energy electric vehicle sales (10,000 units)	Economic factors and costs
2013	1.76	1.344
2014	7.48	1.258
2015	33.11	0.636
2016	50.7	-0.075
2017	77.7	-0.161
2018	125.6	-0.161
2019	120.6	-0.22
2020	136.6	-0.876
2021	192.8	-0.875
2022	210.1	-0.869

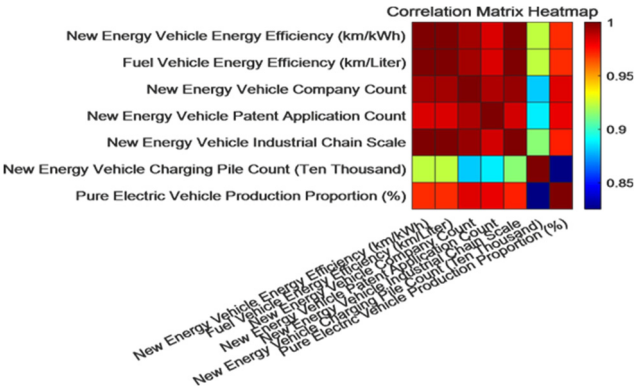

Figure 2. Matrix heat map

Based on the correlation analysis results, we can draw the following conclusions:

Highly correlated indicators: Through the color distribution of the heat map, it can be found that the correlation coefficient between each indicator is relatively large. From the figure, we can find that the correlation coefficients between "fuel vehicle energy efficiency" and "new energy electric vehicle industry chain scale", "new energy electric vehicle companies" and "new energy electric vehicle patent applications" and other indicators are close to 1, which means There is a strong positive correlation between the variables.

From these results, it can be seen that there is a strong positive correlation between the selected indicators. This may indicate that they are interdependent, or that they may be influenced by some common factors. For example, an increase in the number of firms may be associated with an increase in the number of patent applications, which may reflect overall industry growth and increased innovation activity [3].

Therefore, for the infrastructure and innovation category, we directly use the energy efficiency of fuel vehicles (km/liter) as the fourth category indicator to replace other indicators. Finally, the dimensionality reduction results are summarized, and the results are shown in the following Table.4.

Table 4. Dimensionality reduction results

Years	New energy electric vehicle sales (10,000 units)	Economic factors and costs	Social macro indicators	Market and Penetration Category	Infrastructure and Innovation
2013	1.76	1.344	-1.603	-1.118	4.2
2014	7.48	1.258	-1.119	-1.012	4.5
2015	33.11	0.636	-0.72	-0.792	4.8
2016	50.7	-0.075	0.008	-0.601	5.1
2017	77.7	-0.161	0.283	-0.329	5.4
2018	125.6	-0.161	0.441	-0.011	5.7
2019	120.6	-0.22	0.539	0.343	6
2020	136.6	-0.876	0.531	1.073	6.3
2021	192.8	-0.875	0.932	0.934	6.6
2022	210.1	-0.869	0.709	1.515	6.9

According to the results of dimensionality reduction, we can see that the sales volume of new energy electric vehicles: from 2013 to 2022, the growth trend of new energy electric vehicles is obvious, reflecting the increasing market demand. This also reflects the positive effects of new energy policies to a certain extent. From the figure, we can see that the sales volume of new energy electric vehicles declined from 2019 to 2020. At that time, during the most severe period of the epidemic, the market demand for new energy electric vehicles declined, resulting in a decline in sales volume. Special Social conditions will also have an impact on the sales of new energy electric vehicles.

3. Industry Development Forecasts

3.1. Multiple Linear Regression Model

We try to establish a multiple linear regression model and use MATLAB to draw a scatter plot between new energy electric vehicle sales and four first-level indicators from 2013 to 2022, so that we can more intuitively See the relationship, as shown in Figure 3 below:

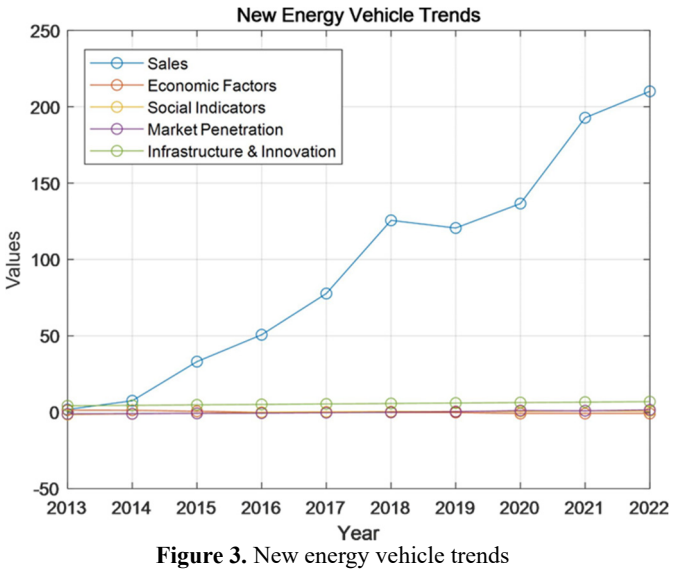

Figure 3. New energy vehicle trends

After analyzing the image and comparing the results of the Person test, we can easily see that the four independent variables have a good linear relationship with the sales of new energy electric vehicles. Therefore, we choose to build a multiple linear regression model as follow [4].

$$\begin{cases} y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 \\ \xi \sim N(0, \sigma^2) \end{cases} \quad (4)$$

Using the regress function in MATLAB for linear prediction, the output result analysis is:

R^2: Indicates the degree to which the fitted model explains the variance of the dependent variable, ranging from 0 to 1. The result is 0.981. Closer to 1 indicates that the model has stronger explanatory power.

Adjust R^2: Adjust the R^2 value based on the number of independent variables in the model. The range is between 0 and 1 and the result is 0.967, a higher result indicates a good model fit.

F statistic: used to test whether the model fit is significant. The result is 66. A higher value indicates that at least one independent variable in the model fits the dependent variable significantly.

p-value: evaluates whether the coefficient of each independent variable in the model is significantly different from zero. The result is 0.000156. A lower result indicates that the model is statistically significant.

Taken together, these indicators show that the linear regression model fits the data well, has high explanatory power and statistical significance. The prediction results are shown in Figure 4.

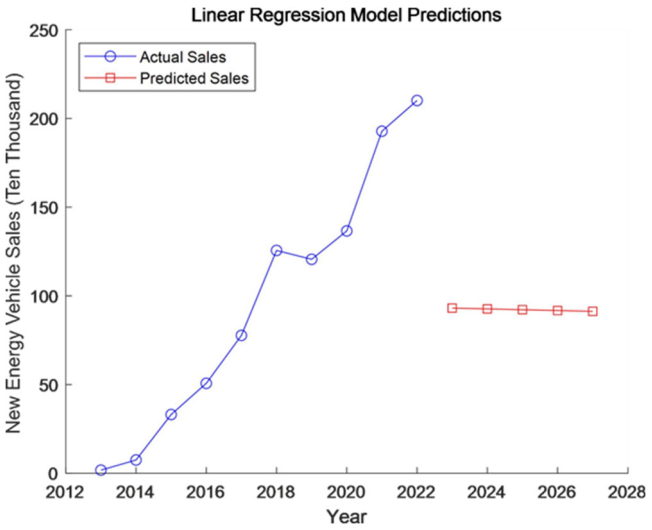

Figure 4. Prediction results

3.2. LSTM Prediction Model

The degree of correlation between each factor and temperature is visualized, and the correlation coefficient is drawn as the above heat map. It can be seen from the figure that the above six influencing factors have high correlation and can be used as predictive temperature factors.

$$x = \frac{X - \min(X)}{\max(X) - \min(X)} \tag{5}$$

Use the data of three core indicators as sample data to determine the number of network input variables. The model diagram is shown in Figure 5 below [5].

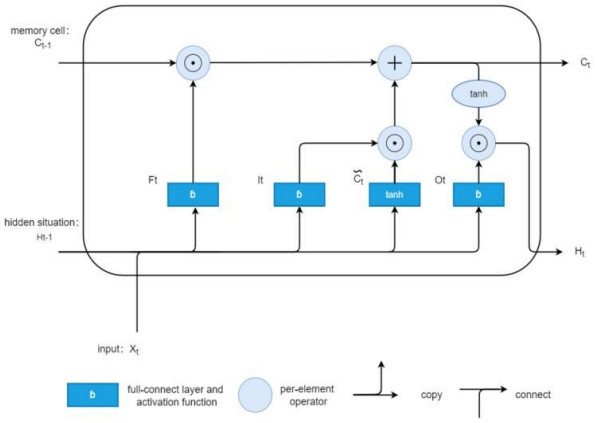

Figure 5. LSTM model

We import the data into MATLAB and use the written program to make predictions. Finally, the prediction results obtained are as shown in Figure 6.

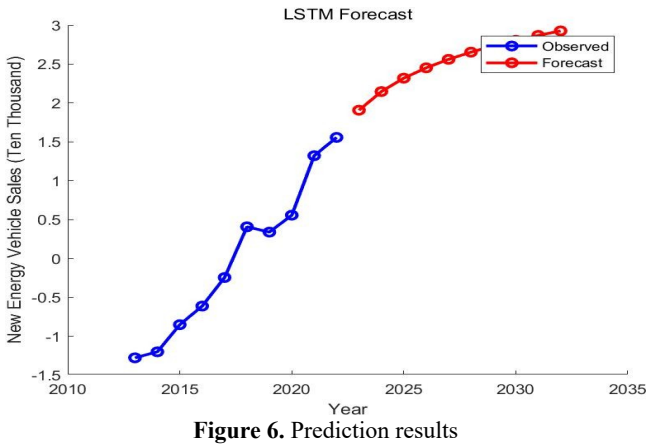

Figure 6. Prediction results

4. Carbon Emission Predictions

The impact of electrification on the ecological environment can be generally considered from the following aspects. The promotion of new energy electric vehicles reduces pollutant gases in the urban environment, which has a direct impact on the health of urban residents and the ecological environment. New energy electric vehicles make less noise than fuel vehicles. The promotion of new energy electric vehicles can reduce noise pollution. The promotion of the use of public transportation such as electric buses can reduce urban traffic congestion and reduce the impact of vehicle exhaust emissions on the environment.

The amount of carbon emissions reflects the quality of the ecological environment to a certain extent. To collect various data from a city with a population of 1 million, we use carbon emissions as the dependent variable, including the number of new energy electric vehicles, sales, market size, The number of enterprises, the number of patent applications, the scale of the industrial chain, the number of charging piles and the coverage of charging infrastructure are used as independent variables. First, consider using a multiple regression model. Analyzing the inherent impact of these indicators on carbon emissions can also reflect the impact on the ecological environment to a certain extent. By constructing a multiple linear regression model, it is found that p value = NaN, F statistic is 0, adjusted R^2 is negative infinity and other abnormal situations, which represents a big problem in using this model. So, we consider using ARIMA to predict future carbon emissions [6]. One model is as follows:

$$F(t) = A + \frac{K - A}{1 + \exp(-r(t - T))} \tag{6}$$

Among them, A represents the initial value, that is, the amount of carbon emissions; K represents the final value, that is, the carbon emissions gradually decrease to zero after reaching the peak; r represents the growth rate, that is, how fast the carbon emissions grow; T represents the peak time, that is, the carbon emissions The time when emissions peak. By fitting historical data and combining the current implementation of various policies and measures, the parameters of the model can be obtained. Predictions are then made based on the model to estimate China's carbon peak and carbon neutrality times. Likewise, we can use ARIMA to predict future carbon emissions. The result is shown in Figure 7.

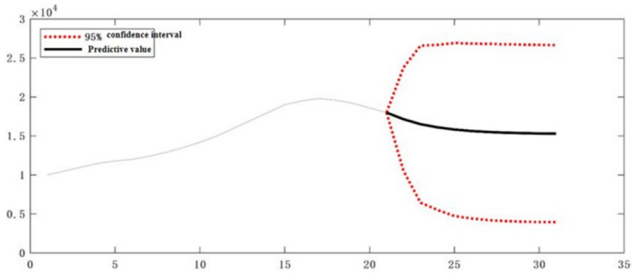

Figure 7. Carbon emission predictions

As can be seen from the figure above, according to our predictions in the next few years, carbon emissions will tend to be in a flat state, but generally speaking, there will be a downward trend in the future. The reduction of carbon emissions can improve the quality of the ecological environment to a certain extent, so it is concluded that the development of electrification can promote the improvement of the environment to a certain extent.

5. Conclusion

This study forecasts the development of China's new energy electric vehicle industry in the next 10 years by using multiple regression prediction model and LSTM model, which provides an important reference for the government and automobile manufacturers. Meanwhile, carbon emissions are predicted by the ARIMA method, revealing the impact of electrification on the ecological environment. We comprehensively analyzed the various factors affecting the development of new energy vehicles and reduced the dimensionality of the data through principal component analysis, which provides a new perspective for an in-depth understanding of the development of the new energy vehicle industry. The application of these methods provides theoretical and practical guidance for promoting the

development of clean energy and sustainable transport and makes a positive contribution to building a cleaner and more environmentally friendly society. In the future, we will continue to pay attention to the development of the new energy vehicle field, constantly improve our research methods, and provide more useful information to promote sustainable transport and ecological environmental protection.

References

[1] Mora-López L, Mora J. An adaptive algorithm for clustering cumulative probability distribution functions using the Kolmogorov–Smirnov two-sample test[J]. Expert Systems with Applications, 2015, 42(8): 4016-4021.

[2] Beattie J R, Esmonde-White F W L. Exploration of principal component analysis: deriving principal component analysis visually using spectra[J]. Applied Spectroscopy, 2021, 75(4): 361-375.

[3] MacDonald J. Electric vehicles to be 35% of global new car sales by 2040[J]. Bloomberg New Energy Finance, 2016, 25(4).

[4] El Aissaoui O, El Alami El Madani Y, Oughdir L, et al. A multiple linear regression-based approach to predict student performance[C]//International conference on advanced intelligent systems for sustainable development. Cham: Springer International Publishing, 2019: 9-23.

[5] Sherstinsky A. Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network[J]. Physica D: Nonlinear Phenomena, 2020, 404: 132306.

[6] Ning L, Pei L, Li F. Forecast of China's carbon emissions based on Arima method[J]. Discrete Dynamics in Nature and Society, 2021, 2021(1): 1441942.