Analyzing New Energy Vehicle Sales Influencing Factors Based on Statistical Learning

Yichen Han

Jinan University-University of Birmingham Joint Institute, Jinan University, Guangzhou, China, 511443

yxc247@student.bham.ac.uk

Abstract. Amid global environmental concerns and the push for sustainable development, the electric vehicle (EV) market is expanding swiftly, recognized as pivotal in curbing greenhouse gas emissions and advancing energy diversification. This thesis investigates factors shaping the EV market and assesses the impact of new energy vehicle development on conventional vehicles. Drawing on industry reports, market data, and relevant literature, it delves into key factors influencing China's new energy vehicle market, such as patent counts, public charging infrastructure, and government subsidies. Based on this analysis, recommendations are proposed to foster the sustainable growth of the new energy vehicle market, including subsidies for related industries, fostering technological innovation, and enhancing charging infrastructure. The study offers insights into the current and future trajectory of the electric vehicle market, revealing rapid expansion at the expense of traditional energy vehicles. By quantifying and visualizing the impact of new energy vehicles on traditional counterparts, this research furnishes valuable insights for policymaking in the new energy vehicle sector.

Keywords: Quantify, Grey Correlation Analysis, Ridge Regression, TOPSIS.

1. Introduction

Over the past decade, the electric vehicle (EV) market has experienced significant growth, not only making breakthroughs in technological development but also gradually becoming an important branch of the automotive industry worldwide. With the increasing awareness of environmental protection and the promotion of clean energy policies, EVs have gained widespread attention due to their zero-emission characteristics [1-4]. However, the sales and popularization of electric vehicles are not only affected by technological innovation and product performance, but also closely related to a variety of factors such as the development of related patents, government subsidies, and the construction of charging infrastructure. In addition, the development of new energy vehicles will also have an impact on traditional automobiles.

In recent years, with the development of new energy vehicles, more and more people have begun to pay attention to the factors affecting their development [1, 2, 4]. Among them, the impact of new energy vehicles on traditional energy vehicles has also received attention from the industry and society [3]. Past studies usually used gray correlation analysis, ridge regression, and correlation analysis to explore the influencing factors, as well as the TOPSIS algorithm to quantify the indicators. For example: Sun et al. assessed the safety index of reservoir dams based on TOPSIS and gray correlation method [5]. Zhong et al. assessed the fire risk of high-rise residential buildings using gray correlation analysis [6]. Zhou et al. used gray correlation analysis and ridge regression to analyze and predict the carbon emissions of cities in Hebei Province [7]. Cheng et al. constructed a cost prediction model for a highway bridge project using a ridge regression optimization algorithm [8]. Geng et al. used ridge regression and LASSO regression methods to analyze the influencing factors of grain yield in Henan Province [9]. However, the application of these methods in studying the development of the new energy automobile industry is relatively small. Considering that the actual situation is affected by a variety of factors, this study synthesizes the advantages of multiple algorithms to analyze the development of new energy vehicles.

The data used for analysis were obtained from various sources, including https://www.iea.org/, https://data.worldbank.org.cn/, https://www.statista.com/, https://www.stats.gov.cn/. We pre-processed the raw data and then performed gray correlation analysis and ridge regression analysis to identify variables that have a significant impact on the sales of new energy vehicles. Subsequently, the TOPSIS algorithm was utilized to score new energy vehicles and conventional vehicles to quantify their development trends. In addition, correlation analysis and ridge regression were used in this thesis to assess the impact of new energy vehicle development on conventional vehicles.

2. Key Factors in New Energy Vehicle Development

2.1. Data Collection and Data Cleaning

The data in this paper is collected from the International Energy Agency, the World Bank, the Chinese National Bureau of Statistics, and other data websites like Statista. To deal with the missing data, we found that some missing data from the dataset of one website could be found in another website's dataset. So, for these kinds of missing data, we carefully compare the data from different sources and fill them in. For the missing values that were not found, we simply used the Polynomial Interpolation to fill them. After that, preprocessing the data, including label coding, removing duplicate values and outliers, filling in missing values, and processing the data columns. Eventually, the data used to analyze the factors of the development of new energy vehicles in China and the impact of the development of new energy vehicles on the traditional energy automobile industry were obtained. The data are shown in table 1.

Table 1. Data for Analysis of the New Energy Vehicle Development

Year	EV. Sales	EV. Produce	Patent	Income	Charge	Battery	Subsidy	Gas. Price
2011	5120	8368	1000	14551	0	138	33877	7.45
2012	9860	12552	1800	16510	0	159	60926	7.85
2013	15730	17533	2024	18311	0	188	88269	8.05
……				……				
2020	1140000	1366000	19739	32189	810000	338	1946000	6.65
2021	3250000	3545000	22043	35128	1150000	349	4602321	5.95
2022	5900000	7058000	24435	36883	1760000	379	5021428	7.41

2.2. Constructing Analytical Models

By considering the influence of different factors on the development of new energy electric vehicles in China, this thesis constructs a discriminative model of the influencing factors of new energy vehicles. Firstly, the index of "sales" is used to judge the development of new energy-electric vehicles. Next, by analyzing the influence of variables on the sales volume, and derive the "main" influencing factors from it. Gray correlation analysis and ridge regression are used to analyze the main influencing factors [6-13].

2.2.1 Gray Correlation Analysis

Gray correlation analysis is a very active branch of gray system theory, the basic idea is based on the sequence of curves of the geometry of the degree of similarity to determine whether the connection between different sequences is close [14].

Gray correlation analysis is a common method for evaluating the correlation between sub-factors and the parent factor, which is the new energy electric vehicle sales in this case. In this study, gray correlation analysis is used to evaluate the correlation between the number of patents (Patent), per capita income (Income), number of public charging piles (Charge), new energy vehicle battery range (Battery), government subsidies (Subsidy) gasoline prices (Gas. Price) and the parent factor new energy electric vehicle sales (EV. Sales).

The basic process of gray correlation is as follows:

Step 1: Determine the characteristic series and the parent series. The comparison sequence is:

$$[\boldsymbol{X}_1' \quad \boldsymbol{X}_2' \quad \cdots \quad \boldsymbol{X}_n'] = \begin{bmatrix} x_1'(1) & x_2'(1) & \cdots & x_n'(1) \\ x_1'(2) & x_2'(2) & \cdots & x_n'(2) \\ \vdots & \vdots & & \vdots \\ x_1'(m) & x_2'(m) & \cdots & x_n'(m) \end{bmatrix} \tag{1}$$

The parent series (i.e., the evaluation criteria) are

$$\boldsymbol{X}_0' = (x_0'(1), x_0'(2), \cdots, x_0'(m))^{\mathrm{T}} \tag{2}$$

Step 2: Standardize the indicator data to reflect the actual situation accurately, excluding differences in units and numerical magnitudes. Normalize the indicators to prevent irrational phenomena.

Step 3: Calculate the correlation coefficient between the elements of each comparison series and the reference series. The resolution coefficient, denoted as ρ, takes the value in (0, 1). A smaller ρ indicates greater differentiation ability among correlation coefficients, with a typical value of 0.5.

Step 4: Calculate the weighted average of the correlation coefficients between each index and the corresponding element of the reference sequence to reflect the correlation between each object and the reference sequence. This is termed the correlation degree and is denoted as:

$$r_{0i} = \frac{1}{m}\sum_{k=1}^{m} W_k \zeta_i(k) \tag{3}$$

Step 5: Analyze the calculation results. According to the size of the gray weighted correlation degree, establish the correlation order of each evaluation object. The larger the correlation degree, the greater the importance of the evaluation object to the evaluation criteria. The final results of the grey correlation degrees between patent, income, charge, battery, subsidiary, and gas. price and sales are shown in Table 2.

Table 2. Grey Correlation Analysis

Patent	Income	Charge	Battery	Subsidy	Gas. Price
0.8604761	0.6353098	0.9738764	0.6536608	0.9098895	0.3897819

The visualization of the standardized variables is shown in Figure 1. The data is from 2011 to 2022.

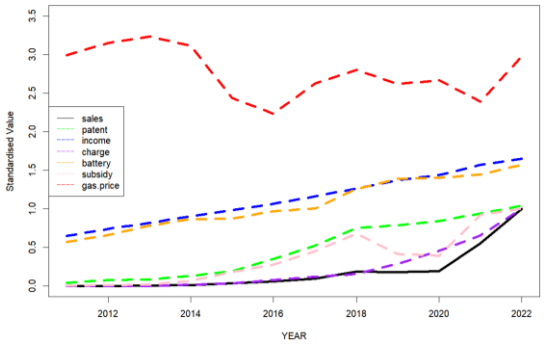

Figure 1. Visualization of Standardized Values

Based on the grey correlation algorithm, it was ultimately determined that public charging piles and subsidy policy exhibit the strongest correlation in the advancement of new energy electric vehicles in China.

2.2.2 Ridge Regression

Ridge regression is a biased estimation regression method dedicated to the analysis of covariate data, which is in essence an improvement of the least squares estimation method. By giving up the unbiasedness of the least squares method, the regression coefficients are more realistic and reliable at the cost of losing part of the information and decreasing the precision, and the fit to pathological data is stronger than that of the least squares method [15].

The principle of determining the K-value through ridge trace plots is the smallest K-value when the standardized regression coefficients of each independent variable tend to stabilize. In general, the smaller the K value, the smaller the bias. The ridge parameter K-value determined by the ridge trace

plot analysis method is subjective and artificial to a certain extent. K = 0.106 was determined automatically using the variance expansion factor method [15].

In the ridge regression, we choose electric vehicle sales (EV. Sales) as the dependent variable, the independent variables are electric vehicle production (EV. Produce), the number of patents (Patent), per capita income (Income), number of public charging piles (Charge), new energy vehicle battery range (Battery), government subsidies (Subsidy) gasoline prices (Gas. Price).

Table 3. Results of Ridge Regression Analysis

K=0.106	Non-Standardized Coefficients	Standardized Coefficients	P	R^2	Adjusted R^2	F
Intercept	-1168859.42	-	0.262			
EV. Produce	0.42	0.491	0.001***			
Patent	-7.187	-0.036	0.5			
Income	5.575	0.023	0.559	0.99	0.973	58.689 (0.001***)
Charge	1.003	0.32	0.004***			
Battery	-405.567	-0.019	0.714			
Subsidy	0.207	0.205	0.033**			
Gas. Price	143475.725	0.066	1.366			

Note: ***, **, and * represent 1%, 5%, and 10% significance levels, respectively.

The results of the analysis are shown in Table 3. The results of the ridge regression show that in the F-test, the level of significance is presented as a p-value of 0.001***, which indicates that it is statistically significant, so we reject the original hypothesis and prove that there is a regression relationship between the independent variables and the dependent variable. In addition, the model has a goodness of fit R^2 of 0.99 and an adjusted R^2 of 0.973, showing an excellent fit of the model.

In addition, when the regression coefficients were analyzed, it was found that the coefficients of the independent variables EV. Produce, Charge, and Subsidy rejected the original hypothesis at the 5% significance level, indicating that their regression coefficients are significantly different from 0 at the 5% significance level. According to common sense, there is a significant covariance between the production and sales of electric vehicles. Since ridge regression can eliminate the effect of multiple covariance, we can ignore the effect of covariance on the model, and thus ignore the effect of electric vehicle production on electric vehicle sales volume. Therefore, it can be concluded that the number of public charging piles and the subsidy policy are important factors affecting the development of EVs in China.

3. Impact of the Development of New Energy Electric Vehicles on Conventional Energy Vehicles

3.1. Data Collection

Table 4. Data for Analysis of the Impact of the Development of New Energy Electric Vehicles on Conventional Energy Vehicles

Year	EV.sales. Share	EV.stock. Share	EV. Δsales	EV. Δstock	Gas. sales. Share	Gas. stock. Share	Gas. Δsales	Gas. Δstock
2011	0.07	0.0074	0.0590	0.0054	99.93	99.9926	0.0036	-0.0054
2012	0.16	0.0200	0.0900	0.0126	99.84	99.9800	-0.0900	-0.0126
2013	0.27	0.0410	0.1100	0.0210	99.73	99.9590	-0.1100	-0.0210
……				……				
2020	4.20	0.8300	1.6000	0.2300	95.80	99.1700	-1.6000	-0.2300
2021	8.70	13000	4.5000	0.4700	91.30	98.7000	-4.5000	-0.4700
2022	14.00	2.1000	5.3000	0.8000	86.00	97.9000	-5.3000	-0.8000

Analyzing the impact of new energy electric vehicles on the conventional energy automotive industry requires a quantitative measure of their impact. We applied the TOPSIS evaluation method to measure the impact and rank the scores. To assess the current status and future potential of the new electric vehicle industry and the traditional fuel vehicle industry, we considered sales rates, inventory rates, and the differences between them over two years. Table 4 shows the dataset used for the analysis.

3.2. Construct the "Influence" Quantified Model

The "Impact" quantitative analysis model utilizes the previously collected data innovatively introduces the TOPSIS algorithm and integrates the previous ridge regression model to validate the results. The TOPSIS method is a commonly used within-group composite evaluation method, which plays an important role in scoring, visualizing, and accurately reflecting the gaps between indicators [5, 16-19]. Ridge regression can eliminate the influence of covariance and determine the regression relationship between independent variables and dependent variables [7-9, 13, 15]. These two methods can quantify the abstract concept of "impact" into figures, and visualize the impact of new energy electric vehicles on the traditional energy automobile industry.

3.2.1 TOPSIS Algorithm

TOPSIS method is a commonly used comprehensive evaluation technique, which can make full use of the raw data and accurately reflect the differences between the evaluation schemes. Its basic process includes processing the normalized data matrix to determine the optimal and the worst options and then calculating the distance between each evaluation object and these options to determine their relative proximity as the basis for evaluating the advantages and disadvantages. This method does not have strict requirements for data distribution and sample size, and the calculation process is simple and easy to implement [16].

3.2.2 Scoring

We normalized the evaluation matrix and calculated the weights of each criterion. First, we used the entropy weight method for model scoring. However, the entropy weight method is based on the information entropy theory, which determines the weights based on the amount of information in the indicators, which may lead to a large difference from the actual situation. Given the large difference between the results of the preliminary run and the actual situation, it is assumed in this problem that the weights of the current situation and future potential are equally important, i.e., the weights of each item are the same.

The TOPSIS algorithm was used to calculate the scores of new energy vehicles (Score. ev) and conventional energy vehicles (Score. gas), as well as to visualize them. The result is shown in table 5:

Table 5. The Score of the TOPSIS Analysis

Year	Score. ev	Rank. ev	Score. gas	Rank. gas
2011	0.311291347	24	0.500779775	3
2012	0.313260093	23	0.498394865	4
2013	0.315169025	22	0.497011978	5
2014	0.318085338	21	0.494898996	6
2015	0.323603396	20	0.490906483	7
2016	0.325375169	19	0.489534017	8
2017	0.336348475	17	0.481452032	9
2018	0.358019552	15	0.46527943	11
2019	0.342134074	16	0.47630147	10
2020	0.383636271	13	0.444451345	12
2021	0.626880773	1	0.373839147	14
2022	0.518581387	2	0.331767855	18

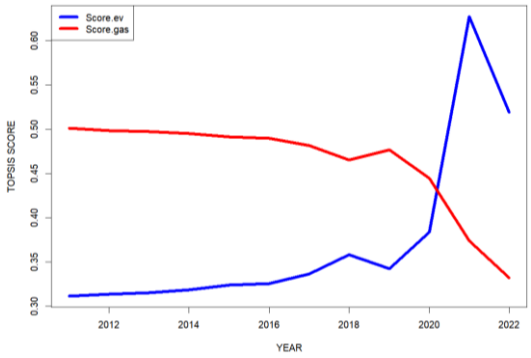

Figure 2. TOPSIS Scores

The left half of Table 5 shows the scores and rankings of new energy vehicles in each year, while the right half represents the scores and rankings of traditional fuel vehicles. As can be seen from Figure 2, since the development of new energy vehicles began in 2011, the situation of traditional energy vehicles has been under threat, and their scores have been declining. Although the decline has slightly stopped from 2018 to 2019, it still cannot stop its continuing downward trend. In addition, looking at the data before 2020, traditional fuel vehicles still maintain a higher score than new energy-electric vehicles. However, after 2020, conventional energy vehicles see a serious decline, while new energy electric vehicles grow significantly, with new energy vehicle ratings overtaking conventional energy vehicles for the first time in our rating model. Although the new energy vehicle ratings fall back in 2022, they are still much higher than those of traditional energy vehicles, and their development continues to be on the upswing. In the past two years, with the support of policy and technology, new energy vehicles have become a strong market force in China.

Next, we analyze the correlation between these two indicators. Spearman's correlation coefficient is used in the analysis as the data in each column grows exponentially over time [20].

Table 6. Results of the Correlation Analysis

	Score. gas	Score. ev
Score. gas	1 (0.000***)	-0.993 (0.000***)
Score. ev	-0.993 (0.000***)	1 (0.000***)

Note: ***, **, and * represent 1%, 5%, and 10% significance levels, respectively.

The results are shown in table 6. Firstly, the correlation coefficient between the scores of traditional energy vehicles and new energy vehicles is less than 0, and the correlation coefficient is not equal to 0 at 1% significance, which can be inferred that there is a significant negative correlation between the two. Secondly, since the absolute value of the Spearman correlation coefficient between the two is close to 1, it indicates that there is a strong correlation between them. In summary, we conclude that the development of new energy vehicles has a significant negative impact on the traditional energy automobile industry.

3.3. Validation of Ridge Regression

Based on the introduction of ridge regression in 2.2.2, we carry out ridge regression on the indicators generated by the quantitative model of "impact". The dependent variable is "Score. gas" and the independent variable is "Score. ev". The K-value was determined from the ridge plot. K=0.001 was automatically determined using the variance expansion factor method.

Table 7. Results of Ridge Regression Analysis

K=0.001	Non-Standardized Coefficients (B)	Standardized Coefficients	P	R^2	Adjusted R^2	F
Intercept	0.65	-	0.000***	0.84	0.824	52.547 (0.000***)
Score. ev	-0.504	-0.916	0.000***			

Note: ***, **, and * represent 1%, 5%, and 10% significance levels, respectively.

The results of the ridge regression are presented in table7: based on the F-test significance p-value of 0.000***, the level of significance, the original hypothesis is rejected, indicating that there is a regression relationship between the independent variable and the dependent variable; In addition, the coefficient of the independent variable "Score. ev" rejects the original hypothesis at the significance level of 1%, and the coefficient of the independent variable is not equal to 0 significantly. At the same time, the model's goodness of fit R^2 is 0.84, adjusted R^2 equal to 0.824, the model fit is more excellent.

The formula of the model: $Score.gas = 0.65 - 0.504 \times Score.ev$

Through the analysis, we found that there is an obvious negative regression relationship between Score. gas and Score. ev, that is to say, we also conclude that the development of new energy vehicles will hurt the traditional energy automobile industry.

4. Conclusion

This paper utilizes gray correlation analysis, ridge regression, correlation analysis, and the TOPSIS scoring model. It begins by investigating crucial factors affecting electric vehicle sales through gray correlation analysis and ridge regression. It identifies patents, charging infrastructure, and government subsidies as pivotal for new energy electric vehicle sales. To foster electric vehicle development, government backing and technological innovation are essential, alongside increased subsidies and faster charging infrastructure construction. The study evaluates current and future prospects of new energy and conventional vehicles based on market sales share, retention rates, and year-on-year changes. Data are scored and compared using the TOPSIS evaluation method, while correlation analysis and ridge regression compare new energy vehicle scores with traditional energy ones, revealing a sustained threat from electric vehicles to conventional markets over 12 years. Although the impact lessened in 2019 before rebounding after 2020, the model's focus is limited to sales and retention of both vehicle types, possibly leading to skewed evaluation weights. Despite potentially less alarming conditions for conventional energy cars, the model assists in identifying trends and assessing present and future scenarios.

This study provides a research framework for analyzing and assessing the new energy vehicle industry. We validate the feasibility of the methodology and highlight the potential applications of our algorithms in industry analysis and assessment. Experimental results show that our model, which combines data mining with grey correlation analysis, ridge regression and TOPSIS algorithms, is able to transcend the limitations of traditional methods and accurately identify the influencing factors of the new energy vehicle industry. In addition, our model is able to quantify the development process of new energy vehicles and conventional vehicles, presenting the results intuitively while establishing correlation and regression relationships between variables. In essence, this study integrates the field of new energy with data science, providing a new perspective for analysis and assessment in related cross-disciplines.

References

[1] Li Chengxin. Analysis of the current situation and development trend of new energy vehicle ownership in China [J]. Automotive Utility Technology, 2024, 49 (04): 8-12.

[2] TANG Baojun, WANG Xiangyu, WANG Bin, etc. Analysis of the development level of China's new energy vehicle industry and its development trend [J]. Analysis and Prospect of China's New Energy Vehicle Industry [J]. Journal of Beijing Institute of Technology (Social Science Edition), 2019, 21 (02): 6-11.

[3] Li Tinghong. Current situation and problems in the development of traditional energy vehicles and new energy vehicles [J]. Science and Technology Innovation Herald, 2013 (06): 67.

[4] Ouyang Minggao. Development strategy and countermeasures of energy-saving and new energy vehicles in China [J]. Automotive Engineering, 2006 (04): 317-321.

[5] Sun Jiachen, Huang Yunchao, Li Songtao. Evaluation of reservoir dam safety index based on TOPSIS and gray correlation method [J]. Science and Technology Innovation and Application, 2024,14 (07): 166-169.

[6] Zhong Yangguan. Fire risk evaluation of high-rise residential buildings based on gray correlation [J]. Journal of the Chinese People's Police University, 2023, 39 (10): 53-57.

[7] ZHOU Lei, GUO Mengjiao. Gray correlation analysis and scenario prediction of urban carbon emissions in Hebei Province [J]. Journal of Hebei University of Science and Technology (Social Science Edition), 2023, 23 (3): 32-42.

[8] CHENG Shu-Xiang, XIONG Shi-Qi, LIU Jun, et al. Cost prediction model of highway bridge project based on ridge regression optimization algorithm [J]. Construction Economy, 2023, 44 (S2): 225-229.

[9] Geng Juan, Nie Wenqian. Analyzing the influencing factors of grain yield in Henan Province based on ridge regression and LASSO regression [J]. Shanxi Agricultural Economics, 2023 (23): 7-10.

[10] Xie Juanjuan. Gray correlation analysis of agricultural industry structure optimization in Gansu Province [J]. Research on Land and Natural Resources, 2024 (02): 41-44.

[11] Ren Zhicheng, Kong Dezhong, Song Gaofeng, et al. Research on prevention and control of general accidents in coal mines based on GRA and AHP [J]. Mining Research and Development, 2023, 43 (12): 131-137.

[12] Luo Yisan. Gray correlation-based monitoring method for river water environment pollution[J]. Resource Conservation and Environmental Protection, 2024 (1): 40-43, 54.

[13] Xing Hong. Analysis on the Influencing Factors and Forecast of Energy Demand in Jiangsu Province under the Background of Carbon Peak [J]. Practice and Understanding of Mathematics, 2024.

[14] TAN Xue-Rui Deng Ju-Long. Gray correlation analysis: a new method for multifactor statistical analysis [J]. Statistical Research, 1995 (03).

[15] LIN Leyi. Application of ridge regression in eliminating multicollinearity [J]. Journal of Liaodong College (Natural Science Edition), 2020, 27 (4): 274-278.

[16] WANG Hui, CHEN Li, CHEN Ken, et al. Multi-indicator comprehensive evaluation method and selection of weighting coefficients [J]. Journal of Guangdong Pharmaceutical University, 2007, 23 (5): 583-589.

[17] JIANG Yue, WANG Xinkai, CHEN Yiying. Evaluation of technological innovation ability of listed enterprises of intelligent manufacturing equipment in China based on entropy weight-TOPSIS method [J]. Henan Science, 2024, 42 (02): 298-305.

[18] Lv Feng, Zhang Shuping, Liu Fen, et al. Evaluation technique of product service system solution based on cloud-TOPSIS-gray correlation analysis model [J]. Modern Manufacturing Engineering, 2024 (02): 38-44.

[19] Xie T, Zhang Yun, Yang HH. Evaluation of foundation pit support program based on combined weights TOPSIS model [J]. Sichuan Construction, 2024, 44 (01): 89-92.

[20] ZHANG Weifeng, XU Weichao. Robustness analysis of Spearman's parsimonious correlation coefficient and Gini gamma correlation coefficient against impulsive noise [J]. Electronic World, 2020 (10): 81-82.