Analysis of the Relationship Between Electricity Supply and Factors in China: Correlation Exploration and Random Forest Modelling

Xiaoyu Liu [1,*,†], Shilin Wang [1,†], Kexin Yang [2], Bing Xie [1]

[1] School of Metallurgy and Resource Engineering, Guilin University of Technology at Nanning, Chongzuo, China, 532100

[2] School of Civil and Surveying Engineering, Guilin University of Technology at Nanning, Chongzuo, China, 532100

* Corresponding Author Email: lxy2397730704@163.com

† These authors contributed equally.

Abstract. This paper focuses on the relationship between China's electricity supply and various factors, and uses data preprocessing, Pearson correlation coefficient analysis, and random forest modelling to conduct the study. Firstly, through the preprocessing and visualisation analysis of the data of multiple indicators, including the selection of nine indicators such as wind power, nuclear power, thermal power and hydropower, etc., the study is carried out. Second, the Pearson correlation coefficient is applied in the correlation factor analysis, and it is pointed out that China's electricity supply is highly correlated with urbanisation degree, coal consumption and other indicators. Subsequently, a random forest model is used for modelling, and the basic process of random forest, the key steps of training decision trees and the analysis of prediction effect are introduced in detail. The final model demonstrates the importance of hydropower in China's electricity development trend, and predicts that China's electricity will be transformed to urbanisation and clean energy in the future.

Keywords: Correlation Analysis, Random Forest, China Power.

1. Introduction

In today's society, with the accelerated industrialisation and urbanisation, the stability and sustainability of electricity supply is crucial for environmental protection and sustainable development. As one of the world's largest energy consumers, the improvement of power supply in China is not only related to the quality of national life, but also to environmental protection. By deeply studying the relationship between power supply and various factors, this paper can better understand the environmental challenges of the power system and find solutions for sustainable development [1]. The aim of this paper is to reveal the potential links between power supply and environmental protection in China through data preprocessing, correlation exploration, and random forest modelling, so as to provide strong support for future environmental protection policy making and power system optimisation. Only by securing power supply can we better protect our common home and achieve the goal of sustainable development [2-3].

2. Model Building and Solving

2.1. Data Preprocessing and Indicator Data Visualisation and Analysis

Data Preprocessing

Obtain data in the open database of the National Bureau of Statistics, the Ministry of Industry and Information Technology, the Energy Bureau and other relevant departments and institutions.

We visualise the relevant indicator data by the frequency of occurrence, and its visualisation results are shown in Figure 1 below:

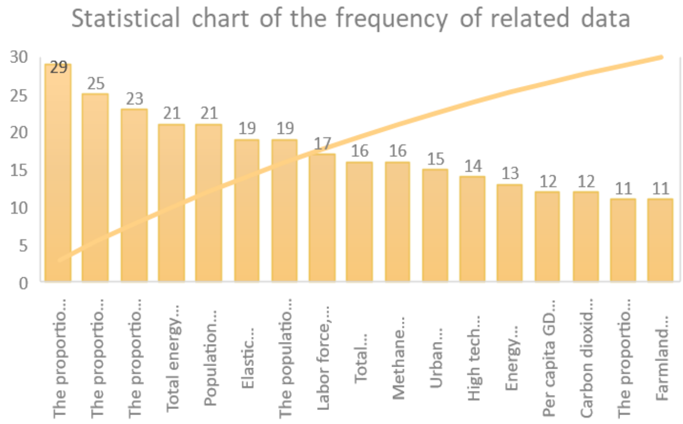

Figure 1. Frequency statistics

Finally, we chose to analyse the nine relevant data on this topic, namely, wind power production, nuclear power production, thermal power production, hydropower production, the proportion of the electricity mix, the value of electricity supply, population growth, the degree of urbanisation, coal consumption, and the level of economic development [4].

Then we find the relevant indicators for data preprocessing, through the 3sigma criterion, in the normal distribution, σ represents the standard deviation, μ represents the mean, x = μ represents the symmetry axis of the image, the 3 σ principle is expressed as shown in Figure 2 below:

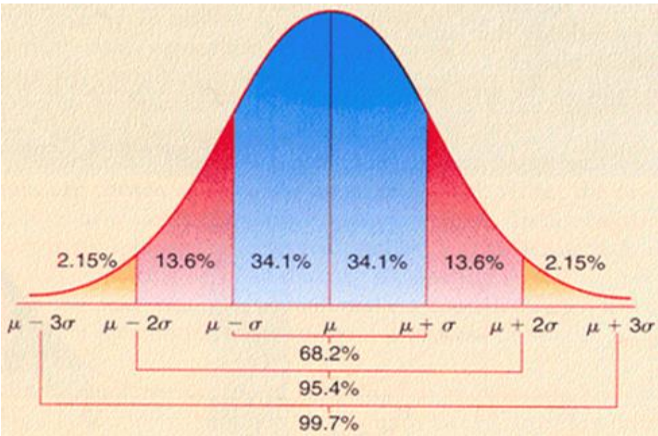

Figure 2. Schematic diagram of 3sigma principle

Indicator data visualisation and analysis

The indicator data visualisation is shown in Figure 3 below:

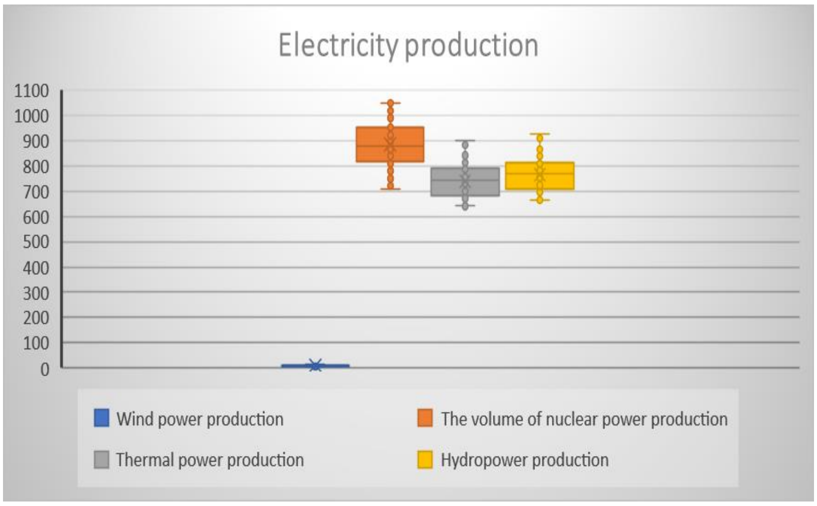

Figure 3. Visualisation of different categories of electricity generation

From the above Figure 3 we can see that the wind power generation is relatively less, now the current main power generation is still thermal, nuclear and hydropower [5-6].

2.2. Correlation factor analysis

Commonly used to test the correlation between the indicators are Pearson correlation coefficient, Spearman correlation coefficient, structural equation modelling, etc. We conducted JB test on the sample data series, and found that there is an absolute normal distribution of the standardized data of the indicator scores of the samples, so we consider using Pearson correlation coefficient for analysis.

Therefore, in this paper, we consider using Pearson's correlation coefficient to explore the correspondence of indicators in the issue of China's electricity supply interacting with many factors, and get relatively reliable results through quantitative analysis [7].

Modelling the Pearson's correlation coefficient

Considering that the Pearson correlation coefficient is used to describe the correlation between two fixed-order data sets, we first sort the size of the indicator scores, transform the quantitative numerical series of this paper into a fixed-order data series, and then calculate the Pearson correlation coefficient between the scores of the indicators, and test the significance of the correlation coefficient. The specific formula for calculating the Pearson correlation coefficient between any two data sets is:

$$p_{X,Y} = \mathrm{corr}(X,Y) = \frac{\mathrm{cav}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y} \tag{1}$$

Which represents the grade difference between the ratings of the two indicators. The interval of change is [-1, 1], with negative values indicating negative correlation and positive values indicating positive correlation, and the closer its absolute value is to 1, the stronger the correlation between the two evaluation indicators. The absolute value identified in this paper is greater than 0.6 for high correlation, less than 0.4 for low correlation, and between the two for medium correlation [8].

Correspondence between indicators

The Pearson correlation coefficient between the scores of the indicators is obtained and calculated as shown in Figure 4 below:

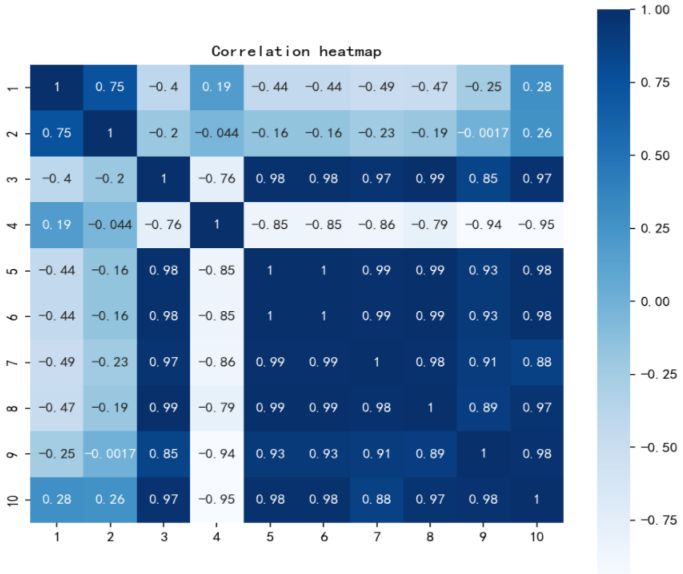

Figure 4. Heat map of Pearson correlation coefficient

Indicators 1-10 in this graph refer to: level of economic development, demographic factors, degree of urbanisation, coal consumption, electricity availability, electricity production, hydropower production, thermal power production, nuclear power production, wind power production, respectively [9-10].

According to the above figure shown in Figure 4 it can be concluded that China's electricity supply is highly correlated with the degree of urbanisation, coal consumption, electricity availability, electricity production, hydropower production, thermal power production, nuclear power production, wind power production, with the specific pattern shown in the above figure.

2.3. Establishment of Random Forest Model

Random Forest is a representative integrated algorithm with all its base evaluators as decision trees. It builds an integrated classifier containing multiple decision trees in a random way, and the forest integrated with regression trees is called the random forest regressor.

The basic flow of Random Forest is shown in Figure 5 below:

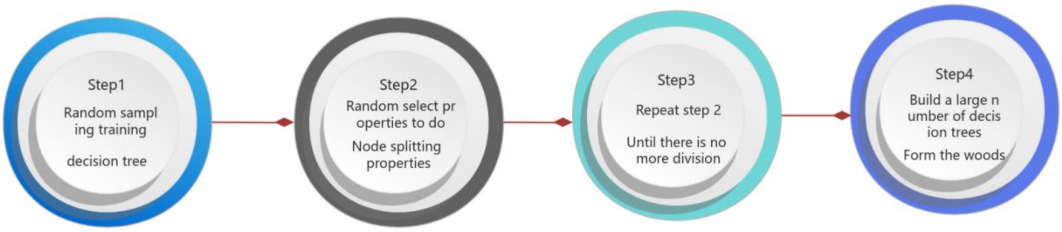

Figure 5. Random forest flowchart

In this step, some hyperparameters need to be set as well as random sampling and feature selection to construct multiple decision trees as a set of classifiers. Specifically, the random sampling method uses Bootstrap sampling method to bootstrap n new datasets from the training set; for each new set of data, m features are randomly selected using the random sampling method; and decision trees are constructed using the generated new datasets and the selected features. At this point we get a forest with multiple decision trees that each independently learn the data and ultimately vote for predictions.

A random forest classification model is essentially an integration of multiple decision trees, each of which can classify data through feature selection and decision rules. Random forests combine the predictions of each decision tree by voting, which leads to the final classification result.

Specifically, if we have n samples and classify them with m features, the probability that the jth sample will be classified in the ith tree is:

$$P_{ij} = \frac{1}{k}\sum_{i=1}^{k}[(x_i, i)] = i \tag{2}$$

The number of decision trees is obtained as shown in Figure 6 below:

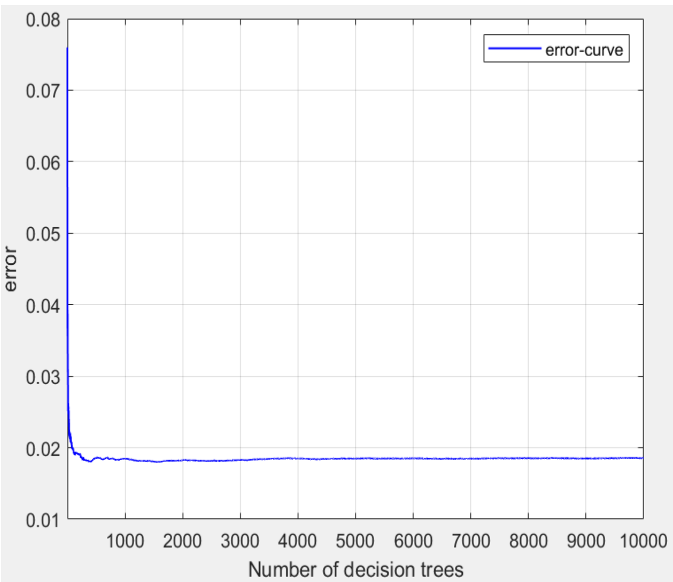

Figure 6. Number of decision trees

$$y_i = argmax\sum_{i=1}^{T} p_i\,(h_i = r_i)(i = 1,\quad 2,\quad \cdots, T) \tag{3}$$

Where T denotes the number of all possible categories and denotes the category to which the tth classifier belongs.

The random forest classification model is implemented by the integration of multiple decision trees. Each tree can be classified by feature selection and decision rules. The overall performance of the model depends on the number of decision trees, hyper-reference settings, and feature selection.

Training Decision Trees

Training Decision Trees produces a decision tree for each dataset. We use the decision tree algorithm CART to train decision trees. The decision tree learning process is divided into two main steps: feature selection and decision tree growth. Feature selection algorithms typically use information entropy or Gini index to measure and gauge the importance of features to determine which features have higher contribution and influence and hopefully make the individual branch subsets purer after division.

Information Entropy (Entropy):

$$H(X) = -\sum_{j=1}^{n} p_i \log_2 p_i \tag{4}$$

Where X denotes a random variable, n is the number of values it takes, and pi denotes the probability that X is equal to i. The higher the information entropy, the more chaotic and unorganised the dataset is.

Gini Index (Gini Index):

$$\mathrm{Gini}(D) = 1 - \sum_{k=1}^{|y|} p_k^2 \tag{5}$$

Where D is the dataset, y is all the categories, and pk is the probability that the kth category occurs in the dataset. The smaller the Gini index, the higher the purity of the dataset.

Decision tree growing divides the dataset into subsets based on the selected features, each corresponding to a particular feature value. For each subset, the above feature selection and decision tree growth steps are repeated recursively until all the samples in all the subsets belong to the same class or the tree reaches a predefined depth or complexity limit.

Firstly, for the current dataset, the information gain or Gini gain needs to be calculated for all features. Secondly, the current feature with the highest gain value is selected as a node and the dataset is divided according to that feature. Perform steps 1 and 2 above recursively for each subset and stop when the leaf node is reached.

Information Gain:

$$Gain(D, A) = H(D) - \sum_{r=1}^{|A|} \frac{D_1|}{D} H(D_r) \tag{6}$$

Where D is the original dataset, A is a feature, $|A|$ is the number of possible values of feature A, and Dv is a subset of instances of a certain value of feature A corresponding to the value of the feature in dataset D. The larger the information gain, the more important the feature is considered to be.

Gini Gain:

$$GainGi(D, A) = Gini(D) - \sum_{N=1}^{A} \frac{Dv}{D} Gini(Dv) \tag{7}$$

Where D is the original dataset, A is a feature, $|A|$ is the number of possible values of feature A, and Dv is the subset of instances corresponding to a certain value of feature A in dataset D. The larger the Gini gain, the more important the feature is considered to be.

In decision tree growth, a common way to select the root node is to select the feature with the largest information gain or the smallest Gini index. For feature selection of non-root nodes, different approaches may need to be considered to balance purity and tree complexity.

Prediction Results Analysis

A comparison of the prediction results of the training set is shown in Figure 7 below:

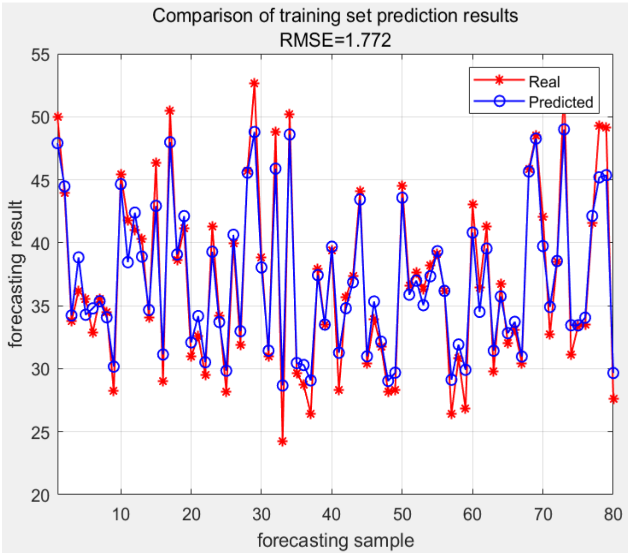

Figure 7. Comparison of prediction results of training set

Generally, we think that there is a better prediction effect in the range of 0-1, but because the data itself has a large RMSE, so in this RMSE=1.772 we still think that this random forest prediction has a better effect.

The importance plot of the indicator data and the comparison of the predicted values vs. the true values of the training set are shown in Figures 8 and 9 below, respectively:

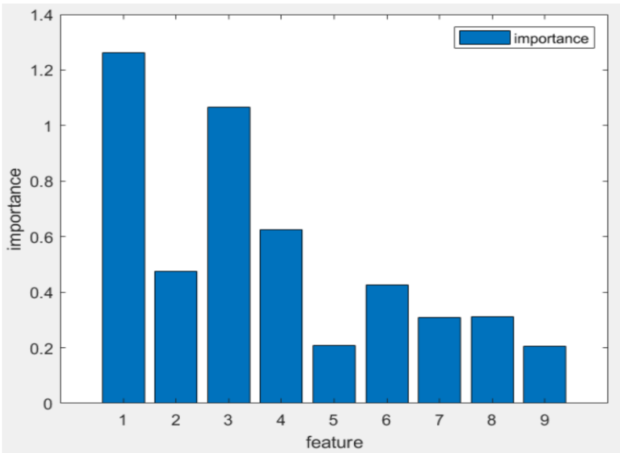

Figure 8. Histogram of the importance of the indicators

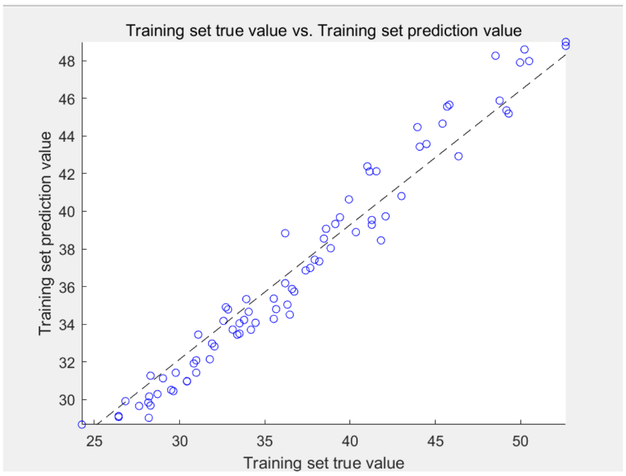

Figure 9. Predicted values vs. real values of the training set

Therefore, we can get that the level of economic development and the degree of urbanisation is the most important, but in terms of the development trend of China's electricity, hydroelectricity is more important, so we can predict that the development trend of China's electricity in 2040-2060 will be in the direction of urbanisation and clean energy.

3. Conclusion

The aim of this paper is to reveal the potential links between power supply and environmental protection in China through data preprocessing, correlation exploration and random forest modelling, which will provide strong support for future environmental protection policy making and power system optimisation. By delving into the relationship between China's power supply and various factors, we can better understand the environmental challenges of the power system and find solutions for sustainable development. By processing and analysing the data of relevant indicators, as well as the establishment of Pearson's correlation coefficient and random forest model, this paper demonstrates the high correlation between China's electric power supply and the degree of urbanisation, coal consumption, and electric power production. The final model predictions show that with the advancement of urbanisation and clean energy, China's future power development will move in a more environmentally sustainable direction. By ensuring a stable power supply, we can better protect our common home and achieve the goal of sustainable development.

References

[1] Wang Xiaoting. New consumption potential in Shenyang accelerates in the first quarter [N]. Shenyang Daily, 2024-05-23 (001).

[2] Zhang Fan. Zhuhai to build new energy industry ecosystem [N]. Zhuhai SAR Daily, 2024-05-22 (001).

[3] PAN Qi, ZHAO Jinquan, WANG Gang, et al. Analysis and calculation method of balance-of-system cost caused by new energy consumption [J/OL]. Power System Automation, 1-9 [2024-05-27].

[4] YANG Shengchun, LU Jianhu, GUO Xiaorui, et al. Modelling and analysis of multivariate flexible resource optimal scheduling adapted to new power system development [J/OL]. Grid Technology, 1-14 [2024-05-27].

[5] Cao Zheng. Beijing to promote new energy vehicle charging new energy electricity [N]. Beijing Urban Vice Centre News, 2024-05-14 (002).

[6] JIN Zhaoli, CHENG Liang. Beijing to promote new energy vehicle charging new energy electricity [N]. Beijing Business News, 2024-05-14 (002).

[7] Cao Xiangkui, Jin Di. "Green content" to enhance the development of "gold content" [N]. China Electricity News, 2024-05-14 (001).

[8] CHENG Qiang, WEI Shuirong, WU Bao, et al. Study on the superposition relationship of expansion force of electric core [J]. Automotive Practical Technology, 2024, 49 (09): 1-5.

[9] Han Zhilin. Renewal of new energy industry for green and high-quality development [N]. Zhangjiakou Daily, 2024-05-13 (001).

[10] Zhao Qiong. Shortcomings remain behind accelerated vehicle-network integration and interaction [N]. China Energy News, 2024-05-13 (008).