Correlation Analysis and Forecast of Power Demand Based on Economic and Meteorological Factors

Junyi Yang1,*, Yi Ge2, Zhenjian Xie2, Junhui Huang3, Hu Li2 and Hui Zhou3

1State Grid Jiangsu Electric Power Co., Ltd, Jiangsu, China
2State Grid Jiangsu Electric Power Co., Ltd. Economic and Technological Research Institute, Jiangsu, China
3Beijing Qingruan Innovation Technology Co., Ltd, Beijing, China

*Corresponding author: 2453305270@qq.com

Abstract. Abstract: Because the economic and meteorological factors are closely related to the level of power demand, this paper establishes an index system for the upstream and downstream industrial chains of important economic and leading industries. The XGBoost algorithm is used to study the quantitative indicators of economic and meteorological factors, and to analyse the time-difference relationship between the economy of the typical city Nanjing, the upstream and downstream industry chain indicators of the leading industry and the power indicators. Through the relevant data of the Nanjing Power Grid, this paper establishes and quantifies a model of the relationship between the economic and industrial chain information and power systems in Nanjing, and predicts the load in the future.

1. Introduction
Power system load forecasting is an important work of the electric power department. It is the premise and basis for power system planning, decision-making and economic operation. Accurate forecasting of power load is of great significance to the safe and economic operation of power system and regional national economic development. There are many factors that affect power load, among which the impact of industrial structure on power demand cannot be ignored [1-3]. The industrial structure plays a leading role in the national economic structure, while the power industry is not only a prerequisite for national economic development, but also a basis for social development and improvement of people's living standards. There is an interactive relationship between power demand and industrial structure. The development of national economy drives changes of industrial structure. Changes of industrial structure cause changes of power demand of various industrial sectors, which affects the development of power industry [4, 5].

Foreign countries began research on power load forecasting as early as the 1950s. The study on power load forecasting in China started in the early 1970s and started later than that in foreign countries for nearly 20 years. The development timeline of power load forecasting at home and abroad is shown in Fig. 1. Since 1980, with the deepening of reform and opening up, the demand for power is extremely strong, and the research on power load forecasting technology has been heating up. Various load characteristic indexes have been defined since 1989 in China, and a systematic analysis of power load characteristics has been carried out. At present, the load characteristic analysis index system of our
country covers 7 daily load characteristic indexes, 9 monthly load characteristic indexes and 14 annual load characteristic indexes [7]. Traditional load forecasting mainly relies on dispatchers or experts to make subjective judgments. Later, with the development of statistical methods, mathematical methods such as time series method, regression analysis method and so on were proposed. With the development of artificial intelligence technology, a series of intelligent algorithms are applied to power load prediction, such as BP neural network, genetic grey RBF model, incremental optimization extreme learning machine, regularized extreme learning machine, support vector machine. Bales and Granger proposed the combination prediction technology in 1969[8], H. Tanaka et al. established a regression analysis prediction model in the 1980s, and Park D.C. established the neural network prediction model for the first time in 1991[2]. The existing load forecasting methods are summarized into two categories: spatial static and time dynamic, and then load forecasting techniques are divided into four categories: traditional forecasting methods, classical forecasting methods, modern forecasting methods and combined forecasting methods [9]. According to the different mathematical models of load prediction, load forecasting techniques are divided into prediction methods based on mathematical probability statistics and prediction methods based on data-driven artificial intelligence by Tang Chenge et al[10].

This paper establishes an index system for the upstream and downstream industrial chains of important economy and leading industries; collects the information of key customers through the negative control system, and analyses the time difference correlation relationship among the economy, the upstream and downstream industry chain indicators of the leading industry and the power indicators of the typical city Nanjing; introduces the relevant theory of econometrics to study the identification method of leading factors, establishes and quantifies a model of the relationship between the economic and industrial chain information and power systems in Nanjing, and predicts the load in the future.

2. Model Theory

Extreme Gradient Boosting (XGBoost) is widely praised by academia and industry because of its fast calculation speed, good model performance, excellent results and efficiency in application practice. XGBoost is used to monitor learning problems and use training data to predict target variables. XGBoost chooses the decision tree as its weak learner. Each time a single weak learner is trained, the weight of the last mistaken data is increased a little before the current single weak learner is trained. Then try to correct the residual error of all the previous weak learners by adding a new weak learner, and finally use the weighted sum of multiple learners together to make the final prediction.

The XGBoost algorithm can be seen as an addition model consisting of K decision trees, as shown in Figure 1.

![Figure 1. Graph of additive model](image)
The formula is shown in formula (1):

\[ \hat{y}_i = \sum_{k=1}^{K} f_k(x_i), f_k \in F \]  

(1)

Where \( f \) is each decision tree; \( F \) is the function space composed of all decision trees.

In the regression process, the parameter \( \Theta = \{ f_1, f_2, \cdots, f_K \} \), so the objective function under the addition model can be changed to:

\[ Obj(\Theta) = \sum_{i=1}^{n} l(\hat{y}_i, y_i) + \sum_{k=1}^{K} \Omega(f_k) \]  

(2)

For the regularization items of decision trees, each decision tree is improved by vector mapping, so the regularization items \( \Omega(f) \) of XGBoost are:

\[ \Omega(f) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^{T} \omega_j^2 \]  

(3)

Where \( T \) is the number of leaf nodes; \( \omega \) is the score vector of the leaf; \( q \) is the function expression that assigns each data point to the leaf.

The complexity is simplified by a forward distribution algorithm. From the beginning to the end, only one base function and its coefficients are learned at each step, and the optimization objective function is approached gradually. In step \( t \), the model predicts \( x_i \) as follows: \( \hat{y}_i = y_i + f_t(x_i) \), where \( f_t(x_i) \) is the decision tree to be learned for this round, so the objective function is:

\[ Obj^{(t)} = \sum_{i=1}^{n} l(\hat{y}_i, y_i) + \sum_{i=1}^{t} \Omega(f_i) \]  

\[ = \sum_{i=1}^{n} l(y_i, y_i + f_t(x_i)) + \sum_{i=1}^{t} \Omega(f_i) \]  

(4)

In this paper, the XGBoost algorithm uses CART tree as the base learning tool. First, the objective function of the load forecasting model is defined, which contains two parts, training loss \( L(\Theta) \) and regularization \( \Omega(\Theta) \), as shown in Formula (5):

\[ Obj(\Theta) = L(\Theta) + \Omega(\Theta) \]  

(5)

Minimizing the objective function means minimizing both the loss function and the regular term. Minimizing the loss function means that the model can better fit the training data well and also means that the model can better fit the real data well. Minimizing the optimization of the regular term reduces the complexity of the model and encourages the algorithm to learn a simpler model. Simple models generally have stable prediction results on test samples and small variances. So optimizing the loss function makes the model out of the under-fitting state, and optimizing the regular term prevents the model from over-fitting. When the whole objective function value reaches a smaller value, the model has a better prediction effect.
Then the loss function $L(\Theta)$ is defined. In this paper, the loss function is commonly used mean squared error (MSE). MSE has first-order terms and quadratic terms, which is more advantageous than other forms of loss function, that is:

$$L(\Theta) = \sum_i (\hat{y}_i - y_i)^2$$  \hspace{1cm} (6)

So the objective function of step $t$ of load forecasting model based on XGBoost is:

$$Obj^{(t)} = \sum_{i=1}^{n} \left( y_i - (\hat{y}_{i}^{t-1} + f_t(x_i)) \right)^2 + \Omega(f_t)$$

$$= \sum_{i=1}^{n} \left[ 2(\hat{y}_i - y_i)f_t(x_i) + f_t(x_i)^2 \right] + \Omega(f_t)$$  \hspace{1cm} (7)

Where $(\hat{y}_i - y_i)$ is the residual.

Make: $g_t = \partial \frac{\partial}{\partial y_1}(y_i - \hat{y}_i)^2 = 2(\hat{y}_i - y_i), h_t = \partial^2 \frac{\partial}{\partial y_1}(y_i - \hat{y}_i)^2 = 2$, bring into the objective function, and combined with the improvement of the regularization term in the previous chapter, make $G_j = \sum_{i \in I_j} g_t$, $H_j = \sum_{i \in I_j} h_t$. At this time, compression results in the following formula:

$$Obj^{(t)} = \sum_{j=1}^{T} \left[ G_j \omega_j + \frac{1}{2}(H_j + \lambda) \omega_j^2 \right] + \gamma T$$  \hspace{1cm} (8)

Assuming that the structure of the XGBoost decision tree is fixed, which means $q(x)$ is fixed, and that the first derivative of $Obj^{(t)}$ is 0, which means the objective function is optimal and can no longer descend along the gradient. Then the corresponding parameter value of the leaf node $j$ can be obtained:

$$\omega_j = -\frac{G_j}{H_j + \lambda}$$  \hspace{1cm} (9)

Therefore, the value of the objective function is:

$$Obj = -\frac{1}{2} \sum_{j=1}^{T} \frac{G_j^2}{H_j + \lambda} + \gamma T$$  \hspace{1cm} (10)

Summary XGBoost prediction model building:

Step 1. A new decision tree is generated in each iteration of the algorithm.

Step 2. Before each iteration starts, calculate the first and second derivatives of the loss function at each training sample point are calculated.

Step 3. A new decision tree is generated by greedy strategies, and the predicted values for each leaf node are calculated by equation (9).
Step 4. Add the newly generated decision tree \( f_i(x) \) to the model: 
\[
y_i = y_i^{t-1} + f_i(x) \tag{10}
\]

Generally, in the fourth step, in order to avoid overfitting the model, the update formula of the model will be changed to: 
\[
y_i = y_i^{t-1} + \alpha f_i(x) \tag{11}
\]
where \( \alpha \) is the step size or learning rate.

3. Organization of the Text

The experimental data comes from the power load in Nanjing High-tech Zone. The forecasting goal is to predict the load condition in the next hour. It is a short-term forecasting task and the forecasting target is the total load in the High-tech Zone. Among them, the data for the whole year of 2016 is used as training data, and the data for the first week of July 2019 is used as test data. The accuracy indicators used in this paper includes root mean square error (RMSE) and mean absolute error (MAE).

\[
\text{MAE} = \frac{1}{N} \sum_{i=1}^{N} |P_i - \overline{P}_i| \tag{11}
\]

\[
\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (P_i - \overline{P}_i)^2} \tag{12}
\]

Where \( N \) is the number of predicted samples; \( P_i \) and \( \overline{P}_i \) are the actual and predicted values at the time \( i \).

3.1. Input Information Settings

The weather data, economic data, grid-side information and calendar rules are used for the input data. As shown in Figure 2, among them, the weather data adopts the maximum temperature, minimum temperature and the average, variance, median, quartile of the wind speed of the month before the forecast date; The calendar rules code statutory holidays (based on Chinese statutory holidays), weekdays, Saturdays and Sundays (some enterprises have different single and double holidays systems); The grid data contains historical load data of each enterprise and the power grid maintenance plan in this area. The economic data refers to the gross industrial output above the scale, emerging industry fixed asset investment, the number of employees and real estate investment related data in the region. In addition, different industrial parks often have different pillar industries. Enterprises in the same high-tech zones tend to produce certain types of products (such as electronic devices, washing products, automotive accessories, biomedicine, new energy power generation equipment, etc.). Therefore, in order to quantitatively analyze the industry dynamics, we quantify the industry stock trends corresponding to the pillar industries in this region as an input factor. In the examples in this paper, 77% of the industries in this high-tech zone are dominated by electrical machinery and equipment manufacturing. Therefore, we take the stock index of the transmission and distribution electric industry as input. The detailed stock data of this industry can refer to reference [25].
As shown in Figure 3, weather and economic characteristics play an important role in the model, in which the average value, variance, and median of the highest temperatures have a greater impact on the model. In the economic data, the industry stock index and industrial output above the scale are also important, which is an important economic indicator causing load changes. In addition, the maintenance plan is closely related to holidays and load changes. Multi-input factor analysis based on XGBoost algorithm can collect important features for model establishment, and effectively assist power demand correlation analysis and prediction impact factor analysis.

**Figure 2.** Setting of input properties.

**Figure 3.** Multi-factor analysis data.
3.2. Forecast Results Analysis

![Figure 4. Load forecasting results.](image)

**Table 1.** Comparison of prediction results for several prediction models.

| prediction model | MAE  | RMSE |
|------------------|------|------|
| linear model     | 32.4 | 28.3 |
| SVM              | 14.2 | 18.5 |
| XGBoost          | 6.6  | 8.11 |

Using the weather and economic related factors obtained in Section 3.1 as the feature input, analyze the advantages and disadvantages of the prediction algorithm in this paper. In order to compare the algorithm with other algorithms and reflect the optimization effect of XGBoost algorithm. The comparison algorithms include linear model and SVM. It can be seen from Table 1 that the average absolute error and root mean square error of the XGBoost short-term load prediction model are both smaller than those of the SVM and linear model prediction models. So this model has higher prediction accuracy and generalization ability. The prediction results are closer to the actual values, and have better prediction results.

4. Conclusion

This paper summarizes the impact of economic and weather factors on load change, analyses the important features that affect load change through XGBoost algorithm, establishes the index system of industry chain upstream and downstream of important economy and leading industries; By collecting key customer information by negative control system, analyze the time difference correlation among economy, upstream and downstream industry chain index of leading industries and power index of the typical city Nanjing; Introduce the relevant theory of econometrics to study the identification method of the leading factor, quantitatively evaluate the model of the relationship among economic, industrial chain information and power system of the typical city Nanjing, and predict the load situation in the future period of Nanjing.

Acknowledgments

This work was financially supported by Theoretical Research on Comprehensive Analysis of Power Economy Based on Big Data Theory, No: J2015039.
References

[1] Boyu Yang, Shijun Chen. Overview of Electrical Load Forecasting Researches and Forecasting Analysis [J]. Sichuan Electric Power Technology, 2018, 41(3):56-60,91. W. Strunk Jr., E.B. White, The Elements of Style, third ed., Macmillan, New York, 1979.

[2] Xianzhong Jian, Yiting Gu. A Short-term Power Load Forecasting Method Based on Full Process Optimization Support Vector Machine [J]. Electric Power Science and Engineering, 2018,34(11):45-51. G.R. Mettam, How to prepare an electronic version of your article, in: B.S. Jones, R.Z. Smith (Eds.), Introduction to the Electronic Age, E-Publishing Inc., New York, 1999, pp. 281-304.

[3] HERNANDEZ L, BALADRON C, AGUIAR J M, etal. A survey on electric power demand forecasting: future trends in smart grids, microgrids and smartbuildings [J]. IEEE Communications Surveys & Tutorials,2014,16(3): 1460-1495. P.G. Clem, M. Rodriguez, J.A. Voigt and C.S. Ashley, U.S. Patent 6,231,666. (2001)

[4] Xiaokan Wang, Qiong Wang. Short-term Load Forecasting Method Based on RBF Adaptive Neural Fuzzy Inference [J]. Journal of South-Central University for Nationalities (Nat. Sci. Edition),2018,37(3):112-115.

[5] Jun Yu. Research on Wireless Sensor Monitoring and Load Forecasting Method for Smart Grid [D]. Taiyuan: Taiyuan University of Technology, 2015.

[6] Shan Li. Electric Power Load Forecasting Method [J]. Beijing: North China Electric Power University, 2017.

[7] Mengkai Wu. Research Analysis and Forecast Method of Medium and Long Term Power Load under New Normal of Economy [D]. Beijing: North China Electric Power University, 2018.

[8] Yanyan Wu. Research on Power Load Forecasting Based on Combination Forecasting Method [D]. Shenyang: Shenyang University of Technology, 2016.

[9] Jingjing Duan. Research on Short-term Load Forecasting Method of Based on Neural Network [D]. Lanzhou: Lanzhou University, 2017.

[10] Chenge Tang. Power load forecasting model based on intelligent optimization algorithm [D]. Nanning: Guangxi University, 2017.