Research on the correlation between electricity consumption and pollutant emission concentration based on DCCA and the prediction of pollutant emission concentration

Guangye Li¹, Jiaxin Zhang¹, Xin Wen¹, Langming Xu¹, Ying Yuan²*  
¹Liaoning Branch of State Grid Corporation State Grid Corporation Shenyang, China  
²School of Business Administration Northeastern University Shenyang, China  
*Corresponding author: yyuan@mail.neu.edu.cn

Abstract. Based on the big data of electricity consumption in industrial enterprises, this paper uses the DCCA (Detrended cross-correlation analysis) method to study the correlation between electricity consumption and pollutant emission concentration. The BP neural network method is also used to model and predict the pollutant emission trends, in order to monitoring of pollutant emissions in industrial enterprises reasonably. The research results show that there is an obvious correlation between electricity consumption data and pollutant concentration data, and the pollutant emission concentration can be predicted to a certain extent by building a BP neural network model. The prediction of pollutants through electricity consumption data helps enterprises to actively respond to the national environmental protection policy and promote national economic development, while allowing more enterprises to enjoy the convenience brought by big data analysis.

Keywords: DCCA, neural network, electric power big data, Associationanalysis, prediction.

1. Introduction
With the rapid development of domestic economic level, the problem of air pollution has become increasingly prominent. The air pollution in China is characterized by high frequency, wide scope and difficult treatment, and the pollutant data that can be obtained is also extremely complex and huge. Therefore, it is important to explore the influencing factors of air pollution and establish an accurate pollution concentration prediction model for air pollution management and promoting the construction of ecological civilization.

According to the data of National Energy Administration, in October 2020, China's industrial electricity consumption accounted for 68.7% of the total national electricity consumption, while thermal power generation accounts for about 70% of the whole equipment in the production of industrial enterprises in China. Thermal power generation is a high energy consumption, high emission and high pollution industry, which pollutes the air environment seriously. The air pollutants emitted in the process of thermal power generation mainly include:SO₂, NOₓ, suspended particulate matter PM and so on. Among them, SO₂ and NO₂ are the main causes of acid rain formation, while suspended particulate matter
PM is the main factor leading to the haze phenomenon. With the rapid and rough development of thermal power generation in China, the air environment pollution in the cities where thermal power generation enterprises are located has increased. And the air environment quality has seriously deteriorated in the areas where thermal power generation is concentrated.

In the context of increasingly stringent national standards for pollutant emission control and the global building of low-carbon cities, it is even more important to effectively grasp and predict pollutant emissions. After understanding the process of thermal power generation, it is not difficult to find that in the production process of thermal power generation enterprises, pollutant generation and emission are often accompanied by the changes in the electricity consumption. Therefore, there may be a certain connection between power consumption and pollutant emissions in industrial enterprises, so it is important to use power data to predict pollutant emissions for enterprises to save energy and reduce emission.

Many scholars at home and abroad have studied on this issue. Wang (2020) used wavelet method to predict pollutant emissions from thermal power plants [1]. Zhang et al (2016), Su et al (2016), and Li and Jiang (2018) obtained better prediction accuracy by improving the prediction method related to pollutants [2-4]. Yang (2017), Lu and Lan (2020), Qiong et al. (2020) used random forest models to predict pollutant concentrations [5-6]. In the production process of industrial enterprises, the generation of pollutants is often accompanied by changes in electricity data, so the prediction of pollutant emissions using electricity data is important for energy conservation and emission reduction of enterprises.

Since the electric power data of industrial enterprises are characterized by many variables, large amounts of data, complex relationship between each variable, and nonlinearity. At the same time, the collection of power data is done by electric meters and other sensors, which makes the collected power data often contain noise, outlier points, and there are certain correlations between the data of each variable, and these characteristics bring great difficulties to the utilization of the data. The scholars at home and abroad have actively explored in this area and achieved certain theoretical results [7-9].

In this context, this paper uses the cutting-edge DCCA method (detrended cross-correlation analysis), collects pollutant emission data and environmental management facilities operation data from thermal power enterprises comprehensively and electricity consumption data from other key emission industrial enterprises. Considering the relationship between industrial production electricity consumption and pollutant emission, conduct correlation analysis between electricity data and key factors of emission. This paper establishes a prediction model for pollutant emission trends according to different industries' emission coefficients of different industries, using deep learning methods to establish a prediction model for pollutant emission trends, and carry out the whole process monitoring of pollutant emissions and environmental management facilities by means of information technology.

This paper is to establish a correlation model between electricity data and pollutants, aiming to help industrial enterprises to predict pollutant emissions by monitoring electricity data in their daily production. It can control pollutant emissions and reduce environmental pollution by providing early warning for their emission targets. It also provides a theoretical basis for the government to formulate reasonable environmental protection policies for different industries according to local conditions. To be specific, this paper firstly uses multidimensional correlation analysis to analyze the correlation between electricity data and pollutant concentrations, in order to filter out the set of variables that have a greater impact on pollutant emissions, which enriches the application of correlation analysis at the industrial level to a certain extent. Further, this paper constructs a pollutant prediction model to predict future pollutant concentrations by monitoring and controlling key electricity data to achieve precise pollution control, further expanding the domestic research on pollutant emission problems. Overall, this study combines theoretical methods with the study of thermal power industry emissions, which both expands the practicality of scientific theoretical methods and provides references for policy makers and researchers in theoretical research and decision planning of thermal power industry.
2. Methodology

Wherever Times is specified, Times Roman or Times New Roman may be used. If neither is available on your word processor, please use the font closest in appearance to Times. Avoid using bit-mapped fonts if possible. True-Type 1 or Open Type fonts are preferred. Please embed symbol fonts, as well, for math, etc.

2.1. Detrended cross-correlation analysis (DCCA)

Detrended cross-correlation analysis (DCCA) is to analyze the nonlinear cross-correlation method between two time series, and compared with the traditional measurement method, the advantage of this method is that it has broken through the original linear paradigm theoretical framework without considering the non-linear and non-normal cross correlation characteristics, and can deeply explore the cross correlation and non-linear complex characteristics between different time series. In addition, it can effectively eliminate the influence of local trend on the scaling behavior of time series, and measure the fractal characteristics of time series in different scales. At present, the DCCA method is widely used to study the cross-correlation between binary non-stationary time series, which can overcome the instability and nonlinear characteristics of the time series and make quantitative analysis. The procedure is as follows:

Step 1: Consider two time series \( \{x_t\} \) and \( \{y_t\} \) of equal length, where \( t=1, 2,\ldots,T \). \( \bar{x} \) and \( \bar{y} \) are the mean of \( \{x_t\} \) and \( \{y_t\} \). Then, we determine the profile as two new series,

\[
X(t) = \sum_{i=1}^{t} (x_i - \bar{x}) \\
Y(t) = \sum_{i=1}^{t} (y_i - \bar{y})
\]

\( t = 1, 2, \ldots, T \).

Step 2: Divide the profiles \( \{X(t)\} \) and \( \{Y(t)\} \) into \( T_s = \text{int}(T/s) \) non-overlapping segments of equal length \( s \). Since the length \( T \) need not be a multiple of the considered time scale \( s \), a short part at the end of the profile will remain in most cases. In order not to disregard this part of the series, the same procedure is repeated starting from the opposite end of the profile series. Thus, \( 2T_s \) segments are obtained together. In this study, we set \( 10 \leq s \leq T/4 \).

Step 3: For each non-overlapping segment \( \nu (1 \leq \nu \leq T_s) \), calculate the local trends \( \{\bar{X}_\nu(t)\} \) and \( \{\bar{Y}_\nu(t)\} \) for each of the \( 2T_s \) segments by least-square fit of each series \( \{X(t)\} \) and \( \{Y(t)\} \) respectively.

Step 4: Calculate the difference between the original time series and the fitting polynomial,

\[
F_\nu^2(s) = \frac{1}{s} \sum_{t=1}^{s} \left[ X_{(\nu-1)s+t}(t) - \bar{X}_\nu(t) \right] \left[ Y_{(\nu-1)s+t}(t) - \bar{Y}_\nu(t) \right]
\]

for \( \nu = 1, \ldots, T_s \) and

\[
F_v^2(s) = \frac{1}{s} \sum_{t=1}^{s} \left[ X_{(\nu-1)s+t}(t) - \bar{X}_\nu(t) \right] \left[ Y_{(\nu-1)s+t}(t) - \bar{Y}_\nu(t) \right]
\]

for \( \nu = T_s + 1, \ldots, 2T_s \).

Step 5: Average over all segments to obtain the cross-correlation fluctuation function,

\[
F_{DCCA}(s) = \left( \frac{1}{2T_s} \sum_{\nu=1}^{2T_s} F_\nu^2(s) \right)^{\frac{1}{2}}
\]

Step 6: Determine the scaling behavior of the fluctuation function by observing log–log plot of \( F_{DCCA}(s) \) versus \( s \), if the original series \( \{x_t\} \) and \( \{y_t\} \) are power-law cross-correlated, as a power-law
where the Hurst exponent $H$ can be obtained by observing the slope of log–log plot of $F_{DCCA}(s)$ versus time scale $s$ via the method of ordinary least squares (OLS). The Hurst exponent $H$ represents the degree of the cross-correlation between the two series $\{x_t\}$ and $\{y_t\}$. $0 < H < 0.5$ indicates anti-persistent (negative) cross-correlation, meaning that if there is an increase of one price, then it is more likely to be followed by a decrease of the other price and vice versa. $H = 0.5$ indicates the absence of cross-correlation, meaning that the change of one price cannot affect the behavior of the other price. $0.5 < H < 1$ indicates persistent long-range cross-correlation (positive), meaning that if one price has been an increase or decrease, then it is likely to be followed by an increase or decrease of the other price, respectively. $1 \leq H < 1.5$ indicates non-stationary yet still mean reverting processes.

Especially, when the time series $\{x_t\}$ is identical to $\{y_t\}$, the detrended covariance $F_{DCCA}(s)$ reduces to the detrended variance $F_{DFA}(s)$,

$$F_{DFA}(s) = \left\{ \frac{1}{2T_s} \sum_{t=1}^{2T_s} \sum_{s=1}^{2T_s} \sum_{s=1}^{2T_s} \right\}^{\frac{1}{2}}$$

### 2.2. BP neural network

BP neural network is a kind of classical learning algorithm. Its main structure is composed of an input layer, several hidden layers and an output layer. Different layers are composed of a number of neurons, and the input information, correlation functions and thresholds determine the output information of each node. Forward propagation of input values and back propagation of errors constitute the learning process of BP neural network. When the forward propagation, the input value along the input layer through the hidden layer, and reach the output layer, after comparing expectations and the computed correlation function of the output value, if the error is bigger, the error of reverse back, before the connection back, after change the weights of each layer of neurons, step by step to reduce the error, such cycles, until the accuracy of the results conform to want.

The specific steps of BP neural network are as follows:

1. Initialization of BP neural network, determine the number of nodes in each layer, and set the initial value of each threshold and weight as a small random number;
2. Input samples and corresponding outputs, learn each sample, and perform steps 3 to 5 for each sample data;
3. Calculate the actual output and the output value of the hidden layer neuron through the input data;
4. Calculate the difference between the expected output value and the actual output value, and calculate the error of the hidden layer and the output layer;
5. Relying on the error calculated in the previous step, replace the connection weights between the relevant nodes of the hidden layer and the input layer and between the nodes of the hidden layer and the output layer;
6. Calculate the error function $F$, and judge whether Converges to the given learning accuracy ($F \leq$ proposed error $E$). If the accuracy requirements are met, the learning ends. If the accuracy requirements are not met, then turn to Step 2 to continue learning.

### 3. Data selection and processing

The data used in this paper are the high-frequency electricity consumption data and pollutant concentration emission data of a cement production line for a cement plant from February 1, 2019 to August 8, 2019. The collected electricity consumption data include electricity meter data of the production line ($X$), the kiln main motor electricity meter data ($Z$). The pollutant data variables include the kiln tail particle (PM) concentration, nitrogen oxide ($NO_x$) concentration and sulfur dioxide ($SO_2$) concentration.
For the existence of missing data and abnormal data, we choose a simple deletion method to deal with missing values and obtain the concentration data of Particulate matter at kiln end (PM), Nitrogen oxides (NOx), Sulfur dioxide (SO2) and original electricity consumption data. Furthermore, we use the box chart method in statistics to identify the outliers of time series data and eliminate them. After eliminating the outlier of these data, we get 201032 observations for PM, 237606 observations for NOx, 110002 observations for SO2. Based on the characteristics of electricity consumption data, we carry out first-order differential processing for electricity consumption data, and then carry out outlier identification and processing. Finally, there are 43622 observations for the electricity consumption data.

4. Correlation between electricity consumption of DCCA and pollutant emission concentration

In order to further explain the correlation between pollutant emission concentration and electricity consumption, DCCA method is used to test the cross correlation between pollutant emission concentration and electricity consumption data. Compared with traditional measurement methods, the advantage of this method is that it breaks through the original linear paradigm theoretical framework without considering the non-linear and non-normal cross correlation characteristics. DCCA method can deeply explore the cross correlation and non-linear complex characteristics between different time series. In addition, it can effectively eliminate the influence of local trend on the scaling behavior of time series, and measure the fractal characteristics of time series in different scales. At present, DCCA method is widely used to study the cross correlation between binary non-stationary time series, which can overcome the instability and nonlinear characteristics of time series and make quantitative analysis. The results of DCCA analysis are shown in Figure 1-6 and table 1. From Table 1, we can draw the following conclusion: the scale index $\lambda$ of pollutant emission concentration and production line electricity consumption is 0.9911, greater than 0.5, close to 1, indicating that there is a strong positive cross correlation between them. That is to say, when the particle concentration has an upward (downward) trend, the power consumption of the production line is likely to show an upward (downward) trend, and the closer the $\lambda$ is to 1, the greater the possibility is. When the scale index $\lambda$ of the particle concentration and the power consumption of the main motor of the kiln is 1.0331, greater than 1, it indicates that the time series is non-stationary, but still has the feature of mean recovery. The scaling index of NOx emission concentration with power consumption of production line and kiln main motor is between 0.82 and 0.88, which indicates that there is a strong positive cross correlation between them. The scaling index $\lambda$ of SO2 emission concentration and power consumption of production line and kiln main motor is between 0.91 and 0.95, which indicates that there is a strong positive cross correlation between them. It is proved that there is a correlation between pollutant emission concentration and electricity consumption data of industrial enterprises.

![Figure 1. DCCA for P-X](image1)

![Figure 2. DCCA for P-Z](image2)
5. Prediction of pollutant concentration based on BP neural network

Based on the correlation analysis results of DCCA method, and considering the influence of the historical concentration of pollutants on the current concentration, a neural network model is established by taking one-period lagged electricity consumption of production line, one-period lagged electricity consumption of kiln main motor, one-period and two-period lagged concentration of pollutants as input variables and the current concentration of pollutants as output variables.

Considering the electricity consumption data, pollutant concentration data, the dimensions of different variables are different, and in order to facilitate the establishment of the model, before the prediction of pollutants, the original data is standardized, and the data is standardized between 0 and 1 according to equation (7).
\[ x^* = \frac{x - x_{\text{min}}}{x_{\text{max}} - x_{\text{min}}} \]  \hspace{2cm} (7)

\[ \text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \]  \hspace{2cm} (8)

In equation (7), \( x \) is the original data; \( x_{\text{max}} \) and \( x_{\text{min}} \) are the corresponding maximum and minimum values.

Secondly, determine the data training set and test set. In this paper, 3 / 4 of the standardized data are randomly selected as the training set, and the remaining 1 / 4 as the test set. The electricity consumption data of the production line, the electricity consumption data of the main motor of the kiln, the one-period and two-period lagged pollutant concentration are taken as the input variables, and the current pollutant concentration data is taken as the output variables to train the BP neural network models of the pollutant concentration prediction. In this paper, the MSE value of equation (8) is used to evaluate the prediction effect of the model. The smaller the MSE value is, the more accurate the prediction is. In equation (8): \( A \) is the total number of samples, \( B \) is the true value, and \( C \) is the predicted value.

\[ x^*(x_{\text{max}} - x_{\text{min}}) + x_{\text{min}} \]  \hspace{2cm} (9)

In this paper, the predicted values of the model are processed by inverse standardization according to equation (9), and compared with the original data of pollutants. Due to the limited space, we only take the kiln tail particulate matter as an example. Figure 7 shows the results of neural network prediction. It can be seen from Figure 7, although the predicted values don't fit the real values so exactly, the overall trend between them is basically similar. Table 2 shows the mean square error of each pollutant concentration prediction based on BP neural network. It can be seen from table 2 that the overall prediction effect is good, and in the three BP neural network prediction models, the prediction accuracy of sulfur dioxide concentration is the most accurate, followed by particulate matter concentration, and the prediction accuracy of nitrogen oxide concentration is the lowest.

| Variable                        | BP neural network |
|---------------------------------|-------------------|
| Particle concentration          | 0.03373464        |
| Nitrogen oxide concentration    | 0.04866629        |
| Sulfur dioxide concentration    | 0.02611381        |

Figure 7. The results of neural network prediction.
6. Conclusions
In the era of Big Data, it is of great practical significance for electricity consumption data mining and pollution control to study the relationship between electricity consumption data and pollutant emission level of industrial enterprises. A large number of electricity consumption data generated by industrial enterprises in the production process can reflect the level of pollutant emissions to a certain extent. In this paper, DCCA method is used to analyze the correlation between the collected electricity consumption data of industrial enterprises and the pollutant concentration data. Based on this cross-correlation analysis, the machine learning method is then used to establish a prediction model for the pollutant emission trend. It indicates that we can apply information means to monitor the pollution management of industrial enterprises.

The empirical results show that there is a strong correlation between the electricity consumption data (including electricity consumption data of the production line and the electricity consumption data of the main motor of the kiln) and the concentration data (including data of particulate matter, nitrogen oxides and sulfur dioxide). According to the association rules, we use the electricity consumption data of the production line, the electricity consumption data of the main motor of the kiln, and the data of one-period and two-period lagged of pollutants to establish the pollutant emission trend prediction model based on BP neural network algorithm. The results of the variance error of the prediction model show that the prediction effect of the model is good.

The research in this paper is one of application of correlation analysis in the field of electric power. Based on electric power data, combining the frontier complexity method with environmental pollution control, we can solve the deficiency of supervision and monitoring in pollutant emission, which can make up for the shortage of traditional pollution control methods. It can also reduce the costs of production and control for enterprises, and realize the construction of ecological civilization with Chinese characteristics and environmental emergencies.

Acknowledgments
This work was supported by Science and technology project of State Grid Corporation of China(2020YF-59).

The Corresponding Author is Yuan Ying.

References
[1] Wang Kefeng. Prediction and Analysis of Total Emissions of Soot and Wastewater Pollutants in Thermal Power Plants [J]. Environmental Science and Management, 2020, 45 (10): 144 - 148.
[2] Zhang Ran, Zhang Shanshan, Shi Yitao, et al. Predictive model of air pollutant emissions from thermal power plants [J]. Chinese Journal of Environmental Engineering, 2016, 10 (5): 2547 - 2550.
[3] Su Yinjiao, Su Tiexiong, Wang Dazhen, et al. Improved wavelet neural network used for prediction of pollutant emissions in thermal power plants [J]. Computer Science, 2016, 43 (1): 508 - 511.
[4] Li Y, Jiang P, She Q, et al. Research on air pollutant concentration prediction method based on self-adaptive neuro-fuzzy weighted extreme learning machine [J]. Environmental Pollution, 2018, 241(10):1115-1127.
[5] Yang S Q. Application Of Random Forest Model In Fine Particle Concentration Prediction In Taiyuan [D]. Taiyuan University of Technology, 2017.
[6] Lu Qiuqin, Lan Qiong, Huang Guangqiu. Research on the prediction of VOCs concentration in spatial correlation region based on grid division [J]. Journal of Xi’an University of Technology, 2020, 11: 1 – 9.
[7] Shen Huijuan, Cao Xiaoli. Research on the application of Apriori association rule algorithm based on frequency set [J]. Internet of Things Technologies, 2020, 10 (10): 57 - 61.
[8] Fang Xin, Yin Jun, Jiang Su, et al. Data Mining Algorithm for Fault Rules of Distribution Network Basedon Combination of Isometric k-means and apriori Algorithm [J]. Smart Power,
2020, 48 (10): 99 - 104+125.

[9] Chen Wenjing. Research on Association Rules of Power Economy Development Based on Apriori Algorithm [J]. Technology and Economic Guide, 2019, 27 (24): 4 - 5.