Research on Correlation Analysis of Power User Behavior Based on Coupled Meteorological Factors

Yixiao Li
School of Electrical Engineering
Shandong University
Jinan, China
202034669@mail.sdu.edu.cn

Tianguang Lv
School of Electrical Engineering
Shandong University
Jinan, China
tlu@sdu.edu.cn

Xin Zhao
Economic & Technology Research Institute of State Grid Shandong Electric Power Company
Jinan, China

Jiyan Liu
State Grid Shandong Electric Power Company
Jinan, China

Wenjie Ju
State Grid Shandong Electric Power Company
Jinan, China

Wanlei Xue
Economic & Technology Research Institute of State Grid Shandong Electric Power Company
Jinan, China

2021 IEEE 4th Student Conference on Electric Machines and Systems (SCEMS) | 978-1-6654-3681-6/21/$31.00 ©2021 IEEE | DOI: 10.1109/SCEMS52239.2021.9646162

***Abstract*—In recent years, large-scale power outages caused by extreme weather conditions have occurred frequently at home and abroad. Hence, how to quickly and accurately use the massive data in the power grid to mine the correlation between meteorological factors and users' consumption behavior is of great significance. Especially in assisting subsequent power dispatching decisions, improving the power grid's ability to respond to extreme weather, and ensuring the reliability of power supply for important power users. In this situation, this paper combine The Random Matrix Theory and Pearson's Correlation Coefficient Theory, to further propose a new method of data mining. Firstly, calculate and select effective coupled meteorological indicators, and then calculate the Pearson correlation coefficient between the indicators and the load data. On this basis, a random matrix model is constructed, and its time-series linear characteristics such as empirical spectrum distribution conform to a variety of statistical theorems. Finally, complete the spectrum analysis, and then fully quantify the degree of coupled meteorological factors on the consumption behavior of users, and realize the effective combination of visualization and quantification. This paper aims to effectively extract the correlation characteristics, and serve for subsequent load forecasting and power system planning. The 2018 and 2019 summer data of a certain regional grid in my country are selected to verify the effectiveness of the proposed method.**

Keywords—random matrix theory, data mining, pearson's correlation coefficient, user behavior, coupled weather indicators.

I. INTRODUCTION

Recently, affected by extreme weather conditions, large-scale power outages have occurred from time to time at home and abroad. The reliability of electricity users cannot be guaranteed, causing huge economic losses. Nowadays, big data characterized with large capacity and high dimension has become an important strategic resource of the power system. Traditional physical modeling can't meet the requirements of real-time status acquisition and load behavior prediction, thus big data technology has a broader prospect.

Random Matrix Theory has attracted wide attention due to its strong data capacity, data fusion and data processing capabilities, high direct utilization of data and flexible algorithms. It can meet the rapid and accurate requirements of data application under emergency conditions, providing strong support for reasonable decision-making. Previously, the random matrix theory has achieved satisfactory results in the fields of big data reconstruction, situation awareness, abnormal data analysis, and fault detection. Literature [1] proposed a micro-grid state awareness method based on the deviation degree of the random matrix spectrum, which aims to solve the status monitoring of large-scale micro-grids. Literature [2] proposed a transmission line fault detection method based on random matrix spectral analysis, which can also complete an identification of weak links. Literature [3] proposed a high-dimensional random matrix for the distribution network failure analysis that mostly relies on its own topology. Theoretical algorithm presented the operating status of the distribution network from the perspective of data and graph theory. Literature [4] applied the random matrix theory to the information extraction of WAMS measurement data, so as to realize the method of determining the fault time and locating the fault area.

At present, research on the correlation analysis between meteorological factors and power systems is mainly aimed at quantitative correlation analysis [5-7] based on single meteorological factors, and qualitative correlation analysis based on coupled meteorological factors. Firstly, for the correlation analysis of a single meteorological factor, quantitative analysis methods such as automatic regression are often used to obtain specific corresponding formulas between the independent variables and the dependent variables. However, the analysis effectiveness is much limited, the data processing algorithm is cumbersome, and the actual versatility

This work was supported by Research and Demonstration of Adjustable Load Marketization Technology under Multiple Trading Mechanism (52060021005S), Science and Technology Project of State Grid Corporation of China.

is poor. It is improper to reflect the coupled effect of multiple meteorological factors under extreme weather conditions. Secondly, in the existing correlation research based on coupled meteorological factors, the validity of indicators based on multiple meteorological factors is not analyzed, and multiple indicators are relatively independent. Literature [5] uses the neural network method to analyze the correlation. This method has certain limitations. Firstly, the weight coefficient that quantifies the correlation cannot directly reflect the degree of correlation. Secondly, in the calculation process, the weight coefficients need to be initialized. If the initialization data is not selected properly, the correlation degree cannot be accurately reflected, and the fluctuations are out of range. Literature [6] established a summer temperature and load effect model, but only describe without quantitatively analyzed, and only put one factor into consideration. Literature [7] does not consider the anti-interference performance of the method and the handling methods for error or error data.

In this situation, this paper proposes a new correlation analysis method based on the combination of Random Matrix Theory and Pearson Correlation Coefficient Theory to quantify the correlation coefficient between multiple coupled meteorological indicators and electricity consumption behavior. Firstly, calculate three typical coupled meteorological indicators based on single meteorological data such as temperature, humidity, and wind speed. Secondly, calculate the Pearson correlation coefficient between the coupled meteorological factor indexes and the corresponding grid load active power data on the same time-scale. Then, use the augmented matrix method and the split window analysis method to model the power system. Finally, complete the correlation analysis between meteorological factors and active load data, and further reveal the influence of meteorological factors on the power consumption behavior of users, and realize the effective combination of visualization and quantification.

II. Coupled Meteorological Indicators Selection

The power grid is susceptible to meteorological factors. The impact of meteorological conditions on the operation of the grid is realized by changing the electricity consumption behavior of power users. A single meteorological factor cannot determine certain fluctuations occurred in the power grid. Therefore, digging the correlation analysis between the coupled meteorological indicators and electricity consumption behavior will provide an important guarantee for subsequent forecasting.

A. Heat Index(HI)

The Heat Index can comprehensively reflect the coupled effect of two single meteorological factors, temperature and relative humidity on the human body's perceived temperature. When the temperature is moderate, the relative humidity has a small effect on the actual temperature perception of the human body. Otherwise, especially in summer and winter, the relative humidity will have a greater impact on the actual temperature perceived by the human body. The *HI* index focuses on the high temperature season and uses the temperature and humidity index to complete the correction of the relative humidity to the temperature index. HI index is calculated by:

$$HI = c_1 + c_2T + c_3R + c_4TR + c_5T^2 + c_6R^2 + c_7T^2R + c_8TR^2 + c_9T^2R^2 \quad (1)$$

c_1=-42.38, c_2=2.049, c_3=10.14, c_4=-0.2248, c_5=-6.838*10^{-3} , c_6=-5.482*10^{-2}, c_7=1.228*10^{-3}, c_8=8.528*10^{-4}, c_9=-1.99*10$^{-6.}$

B. Effective Temperature(ET)

The Effective Temperature refers to a thermal sensation index produced by the human body under different conditions of temperature, humidity and wind speed, and is the manifestation of the coupled effect of three single meteorological factors. In the calculation, the actual temperature is based on the static and saturated atmospheric conditions, which represents different body temperature under different wind speeds, different relative humidity and different temperature conditions. The calculation formula is:

$$T_e = 37 - \frac{(37 - T_a)}{0.68 - 0.14R_h + \frac{1}{1.76} + 1.4V^{0.75}} - 0.29T_a(1 - R_h) \quad (2)$$

C. Comfort Index(CI)

The Comfort Index is a measurement of the coupled effect of three single meteorological factors: temperature, relative humidity and wind speed on the human body, and is used to characterize the degree of human comfort in the atmospheric environment. The calculation formula is listed as follows:

$$k = 1.8T_a - 0.55*(1.8T_a - 26)*(1 - R_h) - 3.2\sqrt{V} + 3.2 \quad (3)$$

III. Pearson Correlation Coefficient Theory

The Pearson Correlation Coefficient, also known as Pearson product-moment correlation coefficient, is a linear correlation coefficient. Pearson's correlation coefficient is a statistical indicator used to reflect the degree of linear correlation between two variables. It focuses more on the relationship between the change trend of one variable and the change trend of another variable, so as to more accurately reflect the relationship between the two variables. The following performance is represented by the symbol r_{pq}, and the value is limited between -1 and 1. The greater the absolute value of r_{pq}, the stronger the correlation is. When r_{pq} is greater than 0, it indicates that the two variables are positively correlated, the two variables have the same changing trend, and the larger the value, the better the following performance. When r_{pq} is less than 0, the two variables are negatively correlated, and the trends of the two variables are opposite. The calculation formula of Pearson coefficient is:

$$r_{pq} = \frac{n\sum_{t=1}^{n} x_p(t)x_q(t) - \sum_{t=1}^{n} x_p(t)\sum_{t=1}^{n} x_q(t)}{\sqrt{\left(n\sum_{t=1}^{n} x_p(t)^2 - \left(\sum_{t=1}^{n} x_p(t)\right)^2\right)^*\left(n\sum_{t=1}^{n} x_q(t)^2 - \left(\sum_{t=1}^{n} x_q(t)\right)^2\right)}} \quad (4)$$

IV. Random Matrix Theory and Data Processing

A. Random Matrix Theory

We define c as the row-column ratio of the random matrix. If the dimension of the matrix tends to infinity and the row-column ratio is fixed, then the time-series linear characteristics of random matrices, such as the empirical spectrum distribution function conforms to various statistical theorems. Therefore, based on time-series data, a random matrix model can be constructed to quickly select and calculate large-capacity data. The analysis method based on the random matrix theory has good robustness, and can be adaptive and stable to phenomena such as erroneous data, data loss, and anomalies.

B. Augemented Matrix Method and Data Processing

According to big data analytics, the relationship between factors and power systems can be revealed by data correlation [3]. This paper uses the augmented matrix method to combine the load active data and the coupled meteorological factors data. We select daily maximum load data and daily average load data of a certain city power grid as the state matrix, which served as the basic part. Basic part describes the state of the power grid. Then we take the coupled meteorological factors data we calculated before as the factor matrix, served as the augmented part, and the above two parts jointly form the augmented matrix as the data source for correlation analysis. Since the number of load data selected in this paper is far greater than the coupled meteorological factors data, the augmented part needs to be expanded. The extension is specifically divided into two steps in order to balance the proportion of factor data and status data. Firstly, we duplicate the factor data for e times. We assume the dimension of state matrix is n_k, then $e = 0.4 * n_k$. Secondly, we introduce white noise into the copied matrix to avoid interference analysis of the correlation due to repeated data. The magnitude of white noise affects the correlation result and needs to be selected carefully. In this paper, we select the signal-to-noise ratio (SNR) ρ as

$$\rho = \frac{Tr(D_e D_e^T)}{Tr(MM^T) * n^2} \tag{5}$$

where Tr(·) is the trace of the matrix. The value of n is calculated by

$$n = \sqrt{\frac{Tr(D_e D_e^T)}{Tr(MM^T) * \rho}} \tag{6}$$

C. Ring Law

We define $\widetilde{X}$ as a standard non-Hermitian random matrix, and its elements are independently and identically distributed variables with

$$\mu(\widetilde{x}_i) = 0 \qquad \sigma^2(\widetilde{x}_i) = 1 \tag{7}$$

Then we regard the matrix product Z as the singular equivalent matrix $\widetilde{X}_u$ of $\widetilde{X}$. We convert the matrix product to the standard matrix product. According to Ring Law, the energy spectral density of the standard matrix product converges almost to the limit with a probability density function defined as

$$f(\lambda_{\widetilde{Z}}) = \frac{1}{\pi c L} |\lambda_{\widetilde{Z}}|^{\frac{2}{L}-2}, \qquad (1-c)\frac{L}{2} \le |\lambda_{\widetilde{Z}}| \le 1 \tag{8}$$

where L is specified as 1 in this paper. The value of inner circle radius is calculated by $(1-c)^{L/2}$, the outer circle radius is a unity on the complex plane of eigenvalues.

V. Case Studies

A. Pearson Correlation Coefficient Results

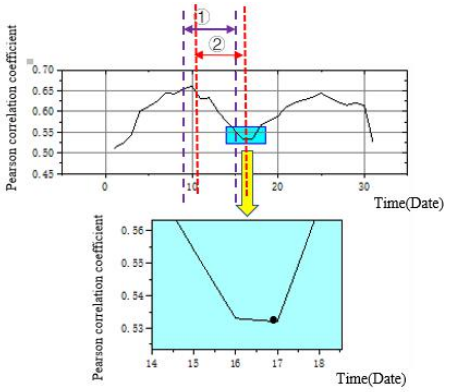

Fig. 1. The Pearson correlation coefficient result of heat index.

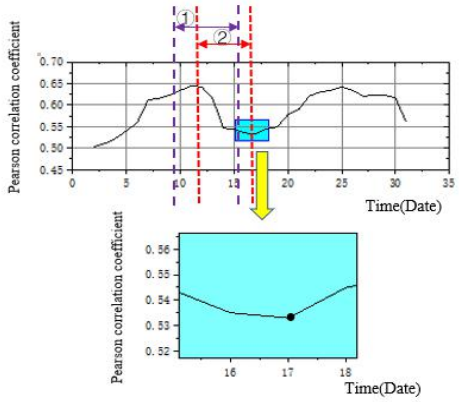

Fig. 2. The Pearson correlation coefficient result of effective temperature.

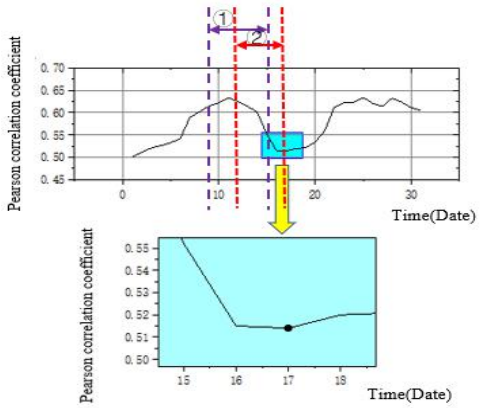

Fig. 3. The Pearson correlation coefficient result of comfort index.

The calculation result of the Pearson coefficient is based on random matrix theory. Take it as a calculation tool for its high-dimension data carrying capacity and high direct availability of data, and also its capacity of rapid selection of the moving window matrix and rapid calculation between the rows of the taken matrix. Here is the background introduction of the selected data: The city ushered in a week of heavy rain starting from August 9th to August 15th, with small-scale fluctuation in temperature, but relative humidity rose more obviously. Starting from August 15th, the temperature continued to rise, and the high temperature conditions continued. The intervals ① shown in the fig.2 to fig.4 indicate the weather conditions actually changed from August 9th to August 15th. The intervals ② shown in the fig.2 to fig.4 indicate the coefficient between each coupled meteorological index and the active load data changes significantly due to the influence of weather conditions.

Analyze from the overall Pearson correlation coefficient curve in August: The change of the correlation coefficient lags behind the change of weather conditions. This is due to the meteorological factors have a cumulative effect on the electricity consumption of power users. The electricity consumption behavior of users does not have a linear statistical characteristic with the coupled meteorological indicators. One of the reasons for choosing random matrix as a tool for data carrying and calculation is cumulative effect. If the data of the active load and meteorological indicators are sampled and calculated every 15 minutes every day, 96-dimensional data is generated in one day, and the Pearson correlation coefficient change caused by the cumulative effect lags no less than 2 days. Then the carry and calculation tools need to carry at least 192-dimensional data. In order to better obtain the following performance of the Pearson coefficient data following the weather coupled index, it should be able to carry at least 384-dimensional data. Therefore, the advantages of random matrix theory in data capacity and fast and accurate calculation can be significantly highlighted.

In the selected interval, the Pearson correlation coefficient shows a downward trend. This is caused by the large-scale start and stop of meteorological loads such as air conditioners. The electricity consumption data declines on a large scale, resulting in the following performance of active load data with the coupled meteorological factors decline significantly. Its' externalization is manifested as a significant decline in Pearson's coefficient.

Analyze from the comparison of the declining rate: the difference in the declining process of the three coupled meteorological indicators is obvious. Compared with the heat index, the comfort index and the effective temperature index have taken more single meteorological indicator into consideration, hence the cumulative effect of meteorological factors is more obvious. The proportion of relative humidity in comfort index and effective temperature is less than the heat index. Therefore, the dropping degree of the comfort index and the effective temperature index is relatively soothing compared with the heat index. The Pearson coefficient of the heat index decreased earlier than the other two coupled meteorological indicators, and the degree of change is more intense.

Analyze from the lowest point of amplitude comparison: the comfort index has the minimum amplitude, the effective temperature index has the maximum. The comfort index changes slowly, and the correlation index value is the smallest. The heat index has the most obvious fluctuation, the correlation index data is in the middle, and the effective temperature index has a smoother decline than the heat index, but the correlation coefficient index is the highest.

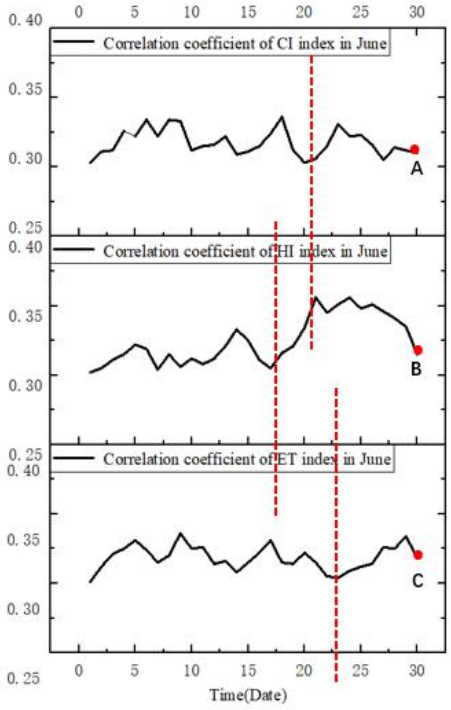

Fig. 4. The Correlation coefficient of three coupled meteorological indicators in June

The overall trend in June did not change much, since June is regarded as a seasonal transition period. The three coupled meteorological indicators have relatively high values, and the human body perceives the climate to be suitable.

Analyze from the overall Pearson correlation coefficient curve in June: Before mid-June: The three coupled indicators have little difference in trend and magnitude. At this time, users' electricity consumption behavior is affected by other social factors such as economy, and less affected by meteorological factors. The correlation values between the three coupled indicators and the user's active power data are all between 0.3 and 0.38, indicating the correlation is weak. In other words, the following performance of users' active power data with meteorological data is poor.

After mid-June: the temperature rises greatly, and the relative humidity does not change much. From fig.5, the

correlation coefficient between the heat index and the user's active electricity data begin to change significantly at the earliest, with the largest increase. Next is the comfort index, and finally is the effective temperature index. Points A, B, and C marked in fig.5 are the correlation coefficients between all the active data in August and the corresponding coupled meteorological index data, which represents the average level of the correlation coefficient index in August. It indicates the effective temperature index has the largest correlation value, followed by the heat index, and finally the comfort index. Therefore, the heat index is the most sensitive index, the effective temperature index has the strongest correlation, and the comfort index performance is relatively poor.

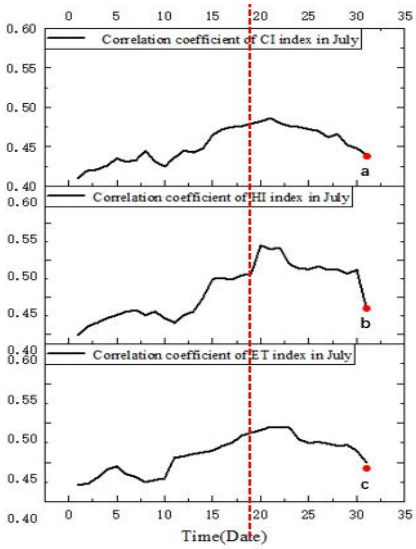

Fig. 5. The Correlation coefficient of three coupled meteorological indicators in July

Analyze from the overall Pearson correlation coefficient curve in July: Before mid-July: The three coupled indicators have little difference in trend and magnitude. At this time, users' electricity consumption behavior is affected by meteorological factors in small scales. The correlation values between the three coupled indicators and the user's active power data are all between 0.4 and 0.495, indicating the correlation is not obvious.

After mid-July: Both the temperature and the relative humidity have risen significantly, and the values of the three coupled meteorological indexes have climbed. The correlation coefficient between the heat index and the user's active power data begin to change significantly at the earliest, with the largest increase. There is little difference between the comfort index and the effective temperature index in the rising range. However, with the relative humidity continue to increase, the values of the three coupled weather indexes doesn't rise but fall, and the user's electricity consumption data continue to rise. Therefore, the correlation coefficients between the three coupled weather indexes and the user's active power data begin to decrease. The heat index dropped at most scales, followed by the effective temperature, and finally is the comfort index. Points a, b, and c marked in fig.6 are the correlation coefficients between all the active data in July and the corresponding coupled meteorological indexes, which represent the average level of the correlation coefficient index in July. It indicates the effective temperature index has the largest correlation value, followed by the heat index, and finally the comfort index. Therefore, the conclusion dragged in July is the same as June.

B. Spectrum Analysis Results of Ring Law

This paper uses the augmented matrix method to combine the load active data and the coupled meteorological factors data. We select daily maximum load data and daily average load data of a certain city power grid as the state matrix, which served as the basic part. Basic part describes the state of the power grid. Then we take the coupled meteorological factors data we calculated before as the factor matrix, served as the augmented part, and the above two parts jointly form the augmented matrix as the data source for correlation analysis.

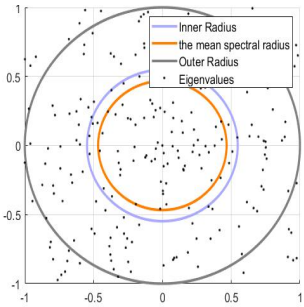

Fig. 6. Eigenvalue distributions of electric power user behavior data in June.

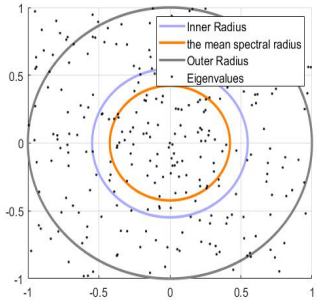

Fig. 7. Eigenvalue distributions of electric power user behavior data in July.

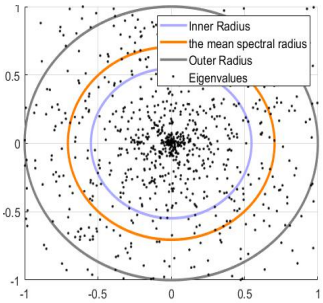

Fig. 8. Eigenvalue distributions of electric power user behavior data in August.

From fig.6 to fig.8, it can be observed that the correlation between coupled meteorological factors and electricity consumption behavior in June and July is obviously lower than that in August. A large number of eigenvalues of the product matrix $\widetilde{Z}$ to be analyzed in August are concentrated inside the inner radius of the circle, which deviates from the predicted ring, revealing that there is a strong correlation between the weather conditions and the electricity consumption behavior of power users.

VI. CONCLUSION

Recently, large-scale power outages caused by extreme weather conditions have occurred frequently at home and abroad. In this situation, this paper proposes a new correlation analysis method based on the combination of Random Matrix Theory and Pearson Correlation Coefficient Theory to quantify the correlation coefficient between multiple coupled meteorological indicators and electricity consumption behavior. Firstly, we calculate three typical coupled meteorological indicators based on several single meteorological indexes, such as temperature, humidity, and wind speed. Secondly, calculate the Pearson correlation coefficient between the coupled meteorological factor indexes and the corresponding grid load active power data on the same time-scale. Then, use the augmented matrix method and the split window analysis method to model the power system. Finally, complete the correlation analysis between meteorological factors and active load data, and further reveal the influence of meteorological factors on the power consumption behavior of users, and realize the effective combination of visualization and quantification.

REFERENCES

[1] Xuguang Hu, Dazhong Ma, Qiuye Sun, Wang Rui.Research on Microgrid State Perception Method Based on Spectral Deviation Degree of Random Matrix [J]. Proceedings of the Chinese Society of Electrical Engineering, 2019, 39(21): 6238-6247.

[2] Xiaoyang Tong, Senlin Yu. Transmission line fault detection algorithm based on random matrix spectrum analysis[J].Power System Automation, 2019, 43(10): 101-108.

[3] Jincan Peng, Blooming Chen, Xiaohua Zhang, Chengming Zhang. Real-time fault analysis method of distribution network based on high-dimensional random matrix theory[J]. Jilin Electric Power, 2021, 49(02):33-38.

[4] Weibiao Chen, Yiping Chen,Yao Wei, Jinyu Wen.A method of fault time determination and fault area location based on random matrix theory[J]. Proceedings of the Chinese Society of Electrical Engineering, 2018, 38(06):1655-1664+1902.

[5] Xipin Zhao, Dai Song, Guoqing Zhang, “Research on correlativity among air temperature, maximum load and power consumption in Shandong power grid,” Power System Technology, 28(17), pp.37-40, Sep.2004.

[6] Hui Zhou, Wenjie Niu, Wantian Liu, “Research on regression model of load effect on specific temperature rise in summer,” Power System Technology, 27(3), pp.46-49, Feb. 2013.

[7] S. Jing, Z. Delv, L. Hu, L. Bingjie, T. Jian and S. Yunli, "Research on Meteorological Sensitive Load in Jiangsu based on Meteorological Dominant Factors," 2019 IEEE Innovative Smart Grid Technologies - Asia (ISGT Asia), 2019, pp.2672-2676.