Application of Correlation Coefficient Method in Identifying Household-Transformer Relationship

Huaqin Qin[1], Yu Han[1], Ziyi Liu[2], Sheng Wang[2], Hongyue Wang[1], Hui Liu[1]

[1] Beijing Kedong Electric Power Control System Co. LTD, 15 Xiaoying East Road, Qinghe, Haidian District, Beijing 100089, China

[2] School of Economics and Management, North China Electric Power University, Beijing 102206, China

Abstract. Since the transformer will produce line loss in the process of distributing power to users, in order to obtain accurate and reliable line loss data, it is very important to identify the correctness of household transformer relationship. The correlation coefficient method can analyze the influence degree of indicators and indicators, indicators and research objects, it has certain advantages in measuring linear correlation and judging the correlation of random variables. Therefore, this paper takes the identification of the relationship between stations and households as the background, and uses the data of power supply, line loss rate and user power consumption as the support to calculate the three correlation coefficients of Pearson, Spearman and Kendall. Based on the calculation results, different stations and time periods are selected to test to verify the reliability of the correlation coefficient method. The test results show that the correlation analysis method combined with three correlation coefficient calculation methods can effectively distinguish whether the household change relationship exists. The technical method of correlation coefficient is more practical for solving the related problems of classification and identification.

Keywords: Station area; Household-transformer relationship; Correlation analysis; Correlation coefficient; Station-user relationship identification.

1. Introduction

The line loss rate of a transformer's power supply range or area is an important economic and technical index of the electric power company. Improving the correct calculable rate of line loss is conducive to reducing losses and increasing efficiency. Since the accurate calculation and analysis of line loss in a station area is based on the correctness of the relationship between user the station area and the household, the identification of the household-transformer relationship becomes more and more meaningful. At the same time, the development of the deepening application of the camp distribution should achieve the sharing of camp and distribution data and information. It also should achieve the coordinated operation of distribution business. Meanwhile, establishing an accurate "substation - line - transformer (station area) - user" correlation, which also puts forward clear requirements for strengthening the grooming of household-transformer relationships and improving the accuracy of household transformation relationship. However, the identification of the relationship between the household and the transformer has always been a major problem for the management of line losses in the power supply enterprises. Although the traditional method of station area outage can check the correspondence between transformers and household meters in the outage area one by one, it is impossible to implement a comprehensive outage check method in order to ensure the reliability of power supply to customers. Other companies use station area identifiers to identify districts by means of carrier communication, but this requires hardware such as concentrators and collectors to be added to the line, and full implementation is limited by cost conditions. For this reason, some studies have added an element of artificial intelligence and combined it with the advantages of big data to analyze the relationship between districts and households. Guochang Li et al [1] studied and experimented with several station area identification algorithms such as clustering and deep learning. It analyzed that using the optimal path method for district identification is a current district identification algorithm with high identification accuracy and low cost; Ya Li et al [2] researched artificial intelligence algorithms based on neural networks for station-home relationship identification;

Alisha Ye et al [3] proposed an intelligent station area identification technique based on spatial-temporal correlation of data; Xiaolin Song [4] proposed a new district-home identification algorithm based on the data collected by smart meters, which can identify the corresponding phase relationship in the station-user relationship with higher identification accuracy rate; Mingming Pan et al [5] proposed a the method of user identification and detection of power theft based on smart meter data, which can effectively improve the intelligence and automation level of power grid enterprises and improve work efficiency; W. Hu et al [6] considered the frequent change of settings in the distribution system and detected the relationship between customers and transformers based on the data collected by smart meters; Xu Huang et al [7] proposed a method for identifying the relationship between station and household based on big data of electricity consumption; L.Ping et al [8] used big data in the power system to achieve online verification of the relationship between households and transformers through a data-driven approach, thus supporting line loss management.

In addition, there are some studies that identify the district-user relationship by profiling the internal relationship between station and user. Jiaju Wang et al [9] proposed the identification method of the station area-user relationship based on multidimensional scale analysis and improved K-means, which could maintain high identification accuracy even with the increase of problem complexity; Topological structure can be used to study the identification of station-user relationship [10-13]; Yuanlin Li et al [14] proposed to take first-order difference stationarity of line loss rate as the indicator for diagnosing abnormal relationship between station and user, and identifying abnormal users in the station by the contribution of line loss fluctuation, so as to further realize local adjustment of household variable relationship; H. Yu et al [15] proposed the intelligent recognition method of grey correlation analysis, which realized the recognition of transformer area in complex environment; The application of hardware and instrument technology is also one of the methods to measure the relationship between user and station [16-18].

To sum up, there are many methods proposed for the identification of station-user relationship at present, but they are basically classified as intelligent algorithms and methods for analyzing internal relationships. This paper analyzes the correlation between line loss and power consumption to identify the wrong users and adjust them. Considering that the use of a correlation coefficient as the basis for judgment is highly subjective, so three major correlation coefficients, Pearson, Spearman and Kendall, are used to identify the relationship between station and household. When the users and the stations calculated by the three correlation coefficients all point to the same, a set of data of the station-household relationship can be output. Reasonable household variable relationship data can improve the correct calculation of line loss in the station and make customers more reliable and transparent. In addition, it can improve service quality and promote loss reduction and energy saving.

2. Basic Theory

The three correlation coefficients are combined to judge the relationship between station and household. The concrete formulae of Pearson correlation coefficient, Spearman correlation coefficient and Kendall correlation coefficient are equation (1), (2) and (3) respectively.

$$\rho_1 = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\left[\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2\right]^{\frac{1}{2}}} \tag{1}$$

$$\rho_2 = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n\left(n^2 - 1\right)} \tag{2}$$

$$\rho_3 = \frac{4P}{n(n-1)} - 1 \tag{3}$$

ρ_1, $\rho_2$and $\rho_3$are Pearson correlation coefficient, Spearman correlation coefficient and Kendall correlation coefficient respectively. Xi and Yi are the i-th number of variables X and Y respectively. $\bar{X}$ and $\bar{Y}$ are the average, and n is the total number of X and Y; d_i refers to the same point i, such as the third X and the third Y. The two groups of variables are simultaneously sorted in order from large to small (or from small to large). The difference between the ranking number of this point in X and the ranking number in Y; P is the logarithm of statistical objects with consistent size relationship between the two attribute values.

For the discrimination of correlation coefficients, there are the following discriminations according to statistics:

Table 1. Correlation coefficient discrimination

Range of correlation coefficients	$\|\rho\|\leq 0.3$	$0.3<\|\rho\|\leq 0.8$	$\|\rho\|\geq 0.8$
Conclusion	No correlation	Weak correlation	Strong correlation

3. Household-transformer relationship identification method

3.1 Principle of identification of household-variable relationship

This paper bases on the DBSCAN clustering algorithm to describe the characteristics of user electricity consumption. And then, the correlation coefficient is obtained according to the correlation analysis technology between users' electricity consumption and the line loss rate of the station area. Next, the correlation coefficient is used to determine whether the household-change relationship exists. The specific steps are as follows:

Firstly, suppose that n+1 data are counted in a certain area, among which there is a statistical error. Assume that the serial number of this user is i*. By definition, $y_t = \sum_{i=1,j\neq i^*}^{n+1} x_i(t) + \varepsilon_t$ (Actual situation); $y_t = \sum_{i=1}^{n+1} x_i(t) + \varepsilon_t'$ (Statistics).

ε_t is the real line loss power consumption, and ε_t' is the statistical line loss power consumption. Then the statistical line loss rate satisfies:

$$k_t' = \frac{\varepsilon_t'}{y_t} = \frac{y_t - \sum_{i=1}^{n+1} x_i(t)}{y_t} = \frac{\varepsilon_t - x_{i^*}(t)}{y_t} = k_t - \frac{x_{i^*}(t)}{y_t} \tag{4}$$

(In formula 4, k_t' is the statistical line loss. $x_i*(t)$ is the electricity consumption from household to user. y_t is the power supply of the station area)

Under the assumption of linear line loss, true line loss rate k_t is a constant. And the statistical line loss rate k_t' will be negatively correlated with $\frac{x_{i^*}(t)}{y_t}$. By calculating the correlation coefficient between the user's electricity consumption and the statistical line loss rate, the user whose correlation coefficient is between [-1, -0.8] can be initially identified as the wrong user who does not belong to the platform area.

On the other hand, for the target platform area, the following equation can also be obtained through similar calculation:

$$y_t = \sum_{i=1}^{n} x_i(t) + \xi_t + x_{i*}(t) \tag{5}$$

$$y_t = \sum_{i=1}^{n} x_i(t) + \xi_t{}' \tag{6}$$

$$\xi_t + x_{i*}(t) = \xi_t{}' \tag{7}$$

So,

$$k_t{}'^* = k_t{}^* + \frac{x_{i*}(t)}{y_t{}^*} \tag{8}$$

yt* refers to the power supply of the target station area on the day t. Then, the correlation coefficient between the user's electricity consumption and the statistical line loss rate of the target station area is calculated. If the correlation coefficient is within [0.8,1], it can be determined that the error of household transformation relationship has occurred. Then, the output data can be further checked manually. The calculation of correlation coefficient mainly includes Pearson correlation coefficient, Spearman correlation coefficient and Kendall correlation coefficient.

3.2 Household-transformer relationship identification process

Based on the above principle of household variable relation identification, the identification process is designed. The general flow chart of calculation and judgment of the station-household correlation coefficient is shown in Figure 1. This process mainly includes data preprocessing module, abnormal station classification module, abnormal user identification module and household change relation adjustment module and result output module. Among them, abnormal station classification module, abnormal user identification module and household change relation adjustment module are the core modules of the algorithm.

(1) Data preprocessing module

This module mainly realizes data reading and preprocessing. The data to be processed includes three parts of data such as household change relation table, table area report scale and user report scale. The function includes reading and calculating the power of the station area and the user. Then get the linked list of to be treated stations. The preprocessing module also checks and processes the missing values, outliers, data consistency and duplicate data.

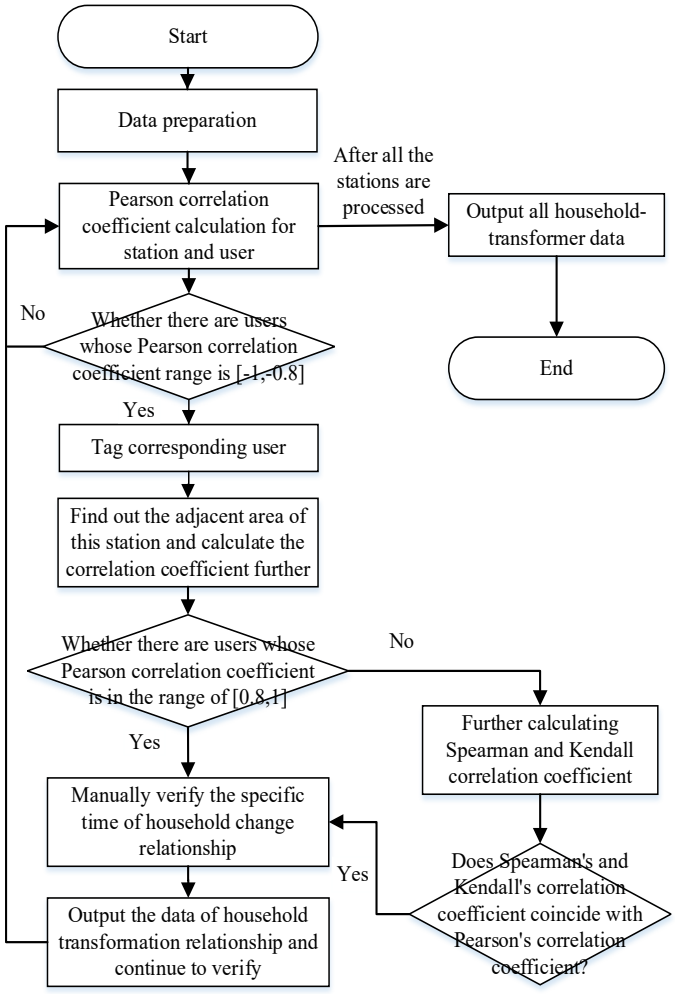


Fig. 1 Overall flow chart of calculation and judgment of station-user correlation coefficient

(2) Pearson correlation coefficient calculation module between station area and user

Based on the line loss rate of the station area and the electricity consumption of the corresponding users, this module calculates Pearson correlation coefficient in a one-month cycle and finds out the users whose correlation coefficient is within the range of [-1, -0.8].

(3) Further verification module between target station area and user

For the user identified in the previous step, the correlation coefficient is calculated with the adjacent station area. If the correlation coefficient is within the range of [0.8,1], it can basically be considered that the user has changed the household transformation relationship.

(4) Inspection module

If there is no station area with correlation coefficient within the range of [0.8,1], Spearman correlation coefficient and Kendall correlation coefficient can be further verified. If the three correlation coefficients are basically consistent in the results and Pearson correlation coefficient is greater than 0.5, it can also be considered that the station-household relationship has changed.

(5) Result output module

According to the data of the user and the corresponding electric quantity of the two stations, the specific time of the user's household variable relationship is verified and the household variable relationship data is output.

4. Example analysis

4.1 Identification of household-transformer relationship

(1) Pearson correlation coefficient test

Pearson correlation coefficient test for users to be tested under the station ID 201069 between October 1 and October 31, 2019. The results are shown in Table 2 (only the first 8 data in ascending order of correlation coefficient are selected, the same below).

Table 2. Pearson correlation coefficient test for different users in the same station area

The user ID	Pearson correlation coefficient
157547335	-0.821284524
158117520	-0.306907095
158094904	-0.220897489
157548295	-0.220064802
158248952	-0.210225836
158103543	-0.194407785
158235330	-0.175805257
157548194	-0.120169464

The Pearson correlation coefficient calculated between the power consumption of user No. 157547335 and the power supply of the station and the line loss rate of the station is close to -1, and there is a large gap with other users. It is preliminarily believed that there is a mistake in the relationship between household and power supply. Considering the household variable relationship test, Pearson correlation coefficient was further calculated for the data from October 1 to October 31, 2019 of all adjacent stations with the ID of 201069. The results are shown in Table 3 (only the first 8 data in descending order of correlation coefficient were selected, the same below).

Table 3. Pearson correlation coefficient test in different stations of the same user

Station area ID	Pearson correlation coefficient
203846	0.544447993
203923	0.485964428
203851	0.460748294
201070	0.425924705
4440257	0.402548543
4074594	0.346996573
201101	0.340336857
201076	0.331236860

It can be seen that the ID of the station with the largest Pearson correlation coefficient is 203846, and the coefficient is 0.54. However, there are still some stations close to the Pearson correlation

coefficient, so the target station cannot be directly determined as 203846, and Spearman correlation coefficient should be further calculated.

(2) Spearman correlation coefficient and Kendall correlation coefficient test

Spearman correlation coefficient test was also conducted between October 1, 2019 and October 31, 2019 for users to be detected in platform area ID 201069. The results are shown in Table 4.

Table 4. Spearman correlation coefficient test for different users in the same station area

The user ID	Spearman correlation coefficient
157547335	-0.770967742
158117520	-0.332258065
157548295	-0.244948974
158094904	-0.222580645
158248952	-0.214919355
158273888	-0.197580645
157547739	-0.157248111
158103543	-0.156854839

It is also found that Spearman correlation coefficient of user number 157547335 is close to -1, and there is a large gap with other users. Further, the results of Table 5 were obtained by testing different stations.

Table 5. Spearman correlation coefficient test in different stations of the same user

Station area ID	Spearman correlation coefficient
203846	0.511290323
203851	0.510887097
203923	0.487500000
201788	0.388143968
201070	0.342275139
203848	0.336459952
201786	0.328258899
201101	0.305474344

The ID of the station with the largest Spearman correlation coefficient is still 203846 and the coefficient is 0.51, but there are still some stations close to it, so the target station cannot be directly determined as 203846. Kendall correlation coefficient is further calculated:

Table 6. Kendall correlation coefficient test for different users in the same station area

The user ID	Kendall correlation coefficient
157547335	-0.595698925
158117520	-0.238709677
157548295	-0.203200203
158094904	-0.161290323
158248952	-0.152688172
157547739	-0.140797894
158273888	-0.122580645
158235330	-0.109677419

It is also found that the Kendall correlation coefficient of user number 157547335 is close to -1, and there is a large gap with other users. The results of Table 7 were obtained by further testing on different stations.

Table 7. Kendall correlation coefficient test in different stations of the same user

Station area ID	Kendall correlation coefficient
203846	0.393548387
203851	0.359139785
203923	0.337634409
201788	0.258342453
201786	0.249731038
4074594	0.240264806
203848	0.237841313
204185	0.222464279

It can be seen that the ID of the station with the largest Kendall correlation coefficient is still 203846 and the coefficient is 0.39.

Through the analysis of the above three correlation coefficients, it can be seen that the absolute value of Pearson correlation coefficient calculated by the power consumption of user number 157547335 and the power supply of the station and the line loss rate of the station is the largest, and the user number with the maximum absolute value of the three correlation coefficients is the same. Meanwhile, the IDs of the stations with the largest correlation coefficients were all the same. Combined with Pearson and Spearman correlation coefficients calculated previously, it can be considered that the following changes in the user relationship have occurred: the user number 157547335 should be adjusted from station 201069 to station 203846.

4.2 Analysis of example results

The correlation coefficient tests conducted by all users 1001~1031 in the station area with ID 201069 from October 1 to October 31, 2019 are drawn as shown in Figure 2. The results of the three types of correlation coefficients are obtained through programming calculation on the experimental platform of MATLAB R2018a.

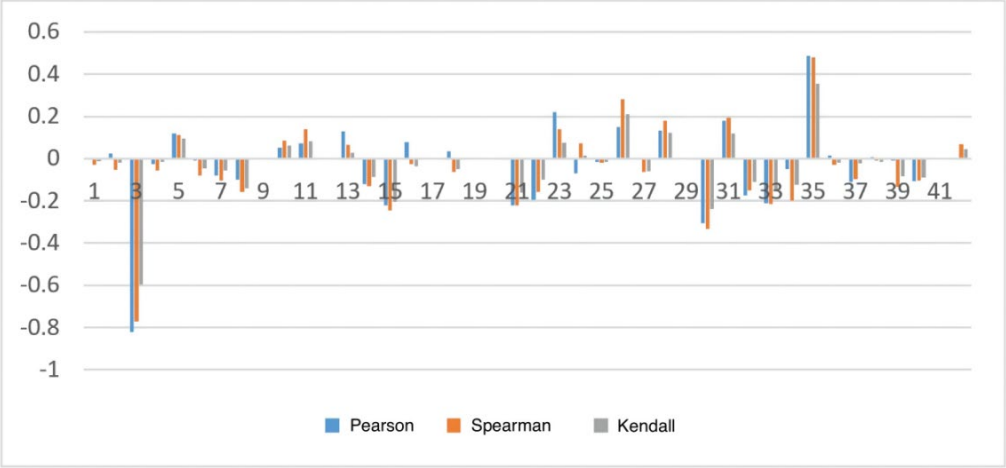


Fig. 2 Inspection of different users in the same station area

It can be seen that the absolute value of correlation coefficient of most users is less than 0.3. That means that there is no correlation. A small number of users show weak correlation, while only one shows strong correlation. This fully demonstrates the feasibility of the correlation coefficient method.

Based on the above calculation, the following error data are obtained. Among them, the first group of output data is the result calculated by the example in this paper. In order to verify that the proposed method is still feasible to identify other stations, another 3 stations were randomly selected for verification. October was still selected as the calculation time of household change for group 2 data,

and December was selected for group 3 and group 4 data. Finally, the output results are summarized to obtain the output data of household variable relationship as shown in Table 9.

Table 8 Output data of household transformer relationship

Group	Original station ID	Target station ID	User ID	Calculation time point of household transform
1	201069	203846	157547335	2019.10.1~2019.10.31
2	4074594	4148882	158460871	2019.10.1~2019.10.31
3	3478123	201692	183447177	2019.12.1~2019.12.31
4	201692	5484531	174550093	2019.12.1~2019.12.31

Based on the data in Table 9, the user number should be adjusted from the original station to the target station. After that, the specific time point of household transform can be obtained through specific verification, and the relationship between station area and households can be adjusted. Obviously, the analysis using three correlation coefficients at the same time is more convincing than that using a single correlation coefficient, which can enhance the reliability of station-household relationship identification to a certain extent. However, there are still some subjective factors involved, and there is a deviation from the actual situation in the identification of multiple users' household change in a station area. Further research and discussion are needed on how to further improve the detection accuracy.

5. Conclusion

Based on the importance and necessity of the identification of station-user relationship, this paper proposes a new method to determine whether there are anomalies in station-household relationship based on the calculation of three correlation coefficients. And it makes an empirical study on this method. The conclusions are as follows: (1) The accurate and reliable test method of customer relationship can improve the work efficiency, and it has guiding function to the analysis of line loss. (2) It is limited to determine the target region directly by only one verification method of correlation coefficient calculation. While it is more reliable to combine Pearson's, Spearman's and Kendall's correlation coefficient analysis. (3) The three correlation coefficient calculation methods are combined to determine whether there is a household-transformer relationship, which provides theoretical guidance for further application of data mining to identify the station-user relationship. This method improves the identification efficiency of the station-user relationship and has high feasibility in practical engineering applications.

References

[1] Guochang Li, Weiqiong Song, Huizhu Xian, Liu Han, Xiao Ye, Juan Yuan. Research and application of power meter station recognition method based on Zero crossing time and SNR Algorithm [J]. Electrical measurement and instrumentation, 2019, 56(07):148-152.

[2] Ya Li, Rusen Fan,Wei Jiang, Junjie Yang, Tao Song, Qinxue Zhao. Research on Intelligent station area identification method based on BP neural network [J]. Electrical measurement and instrumentation, 2017, 54 (03): 25-30.

[3] Alisha Ye, Jun Gu, Xiaoqiu Zhang. Intelligent station area identification technology based on spatial-temporal correlation of data [J]. Modern electronic technology, 2018, 41 (04): 30-33.

[4] Xiaolin Song, Luhan Huang, Yunlong He, Yuanfeng Zhang, Lijian Chen, Yue Chen. A new method for identifying the relationship between satation area and households based on the data collected by smart meters * [J / OL]. Electrical measurement and instrumentation: 1-7.

[5] Mingming Pan, Shiming Tian, Bo Wu, Jun Ye. Research on area identification and power stealing detection method based on smart meter data [J]. Smart power, 2017, 45 (12): 80-84.

[6] W. Hu, Y. Liu, Q. Guo, J. Wang, Y. Wang and Z. Zhao, "Detection of users-transformer relationship in the secondary power distribution system with smart meter data," 2020 15th IEEE Conference on Industrial Electronics and Applications (ICIEA), Kristiansand, Norway, 2020, pp. 449-454, doi: 10.1109/ICIEA48937.2020.9248159.

[7] Xu Huang, Weiheng Wang, Shuang Wu, Wei Hu. Research on identification method of district-household relationship based on power consumption big data [J]. Power supply and consumption, 2019, 36 (10): 22-29.

[8] L.Ping et al. "The Research of Users-transformer Relationship Verification Method Based on Data-driven," 2019 IEEE Innovative Smart Grid Technologies - Asia (ISGT Asia), Chengdu, China, 2019, pp. 2550-2554, doi: 10.1109/ISGT-Asia. 2019.8881160.

[9] Jiaju Wang, Zhongbing Wan, Zhongxiao He, Jia Wang, Zhi Xie, Xiao Wang. Identification method of station-household relationship based on multidimensional scale analysis and improved k-means [J]. Electrical automation, 2020, 42 (02): 56-59.

[10] Zikuan Lian. Identification of topological relations of low voltage stations based on data driven [D]. Zhejiang University, 2020.

[11] Zepu Gao, Yun Zhao, Yilan Yu, Yongjian Luo, Ziwei Xu, Lianmei Zhang. Low voltage distribution network topology identification method based on knowledge map [J]. Power system protection and control, 2020, 48 (02): 34-43.

[12] Guochang Li. Topology identification method based on Lora technology and GPU acceleration [J]. Electrical measurement and instrumentation, 2019, 56 (21): 88-95.

[13] Yong Xiao, Yun Zhao, Zhidong Tu, Bin Qian, Runmian Chang. Low voltage distribution network topology verification method based on improved Pearson correlation coefficient [J]. Power system protection and control, 2019, 47 (11): 37-43.

[14] Yuanlin Li, Fan Deng, Jie Gong, Kang Fu. Identification and adjustment method of household variable relationship considering spatial constraints [J]. Computer knowledge and technology, 2018, 14 (28): 191-194 + 196.

[15] H. Yu, S. Xia, W. Wei, B. Sun, C. Dong, X. He. "Improvement on Application of Gray Correlation Analysis in Intelligent Transformer Area Recognition," 2018 2nd IEEE Conference on Energy Internet and Energy System Integration (EI2), Beijing, 2018, pp. 1-4, doi: 10.1109/EI2.2018.8582387.

[16] Chao Yuan. Research and development of low voltage distribution network area identification technology [D]. Dalian University of technology, 2014.

[17] Di Fan, Shaoshan Li, Haitao Li, Wei Wang. Research on the application of key technology of user identifier in station area [J]. North China power technology, 2010 (07): 27-30.

[18] Rui Yang. Research on user network identification technology in distribution area [D]. Xi'an University of technology, 2018.