# Electricity Anomaly Detection Research of Flue-Cured Tobacco Users Considering the Characteristics of Industry Electricity Consumption Behavior

Xu Zhengyi[1(✉)], Yang Jianping[1], and Liang Yunhua[2]

[1] Safety Supervision Center of State Grid, Hunan Electric Power Co., Ltd., Changsha 410004, China
1182667599@qq.com

[2] State Grid Hunan Electric Power Company Extra High Voltage Substation Company, Changsha 410004, China

**Abstract.** Power theft inspection is an effective means to improve the operating income of power supply enterprises. However, in practice, there has been a problem that the accuracy rate of power theft detection is not high. Cluster analysis of power consumption behaviors by subdividing power users by industry will not only make it possible to extract characteristic index items that accurately describe user behaviors in combination with industry power consumption behavior characteristics, but also reduce user classification dimensions and false positives for power theft detection. Rate is the development direction of promoting the practical application of data-driven power theft detection. This article takes flue-cured tobacco users as the object. Firstly, it analyzes the industry electricity consumption patterns of flue-cured tobacco users, and then based on the statistical analysis of industry characteristics, establishes power consumption index characteristic items that accurately describe the industry characteristics of flue-cured tobacco users. The propagation clustering algorithm performs cluster analysis of users in flue-cured tobacco industry. Numerical simulation results show that the proposed method can not only accurately identify misclassified non-flue-cured tobacco users such as public properties in rural areas, reduce the scope of power theft detection, but also effectively detect specific types of flue-cured tobacco specific power stealing behavior.

**Keywords:** Industry Electricity Characteristics · Electricity Theft Detection · Tobacco Load · AP Clustering

## 1 Introduction

At present, The State Grid Corporation of China have achieved complete collection of electricity consumption information, and can basically grasp the user's electricity consumption data and customer information timely and accurately [1]. On this basis,


H. Yang et al. (Eds.): SmartGIFT 2022, LNICST 483, pp. 203–217, 2023.
https://doi.org/10.1007/978-3-031-31733-0_18

the production technicians summarized the indicators of exact physical meaning such as zero-sequence current, power reversal and voltage loss of electric meters for low-voltage users based on their experience, which can accurately identify abnormal electricity consumption behaviors. However, this type of method is mainly applicable to specific stealing methods, and the coverage of stealing methods is limited. Electric power scientists and technicians have carried out a large number of data-driven electricity anomaly detection research mainly from two aspects: unsupervised cluster analysis and supervised classification analysis, focusing on massive electricity metering data. Among them, unsupervised anomaly recognition is generally based on user data to calculate characteristic index items and then perform cluster analysis, identify abnormal users according to category [2]. Supervised electricity theft detection takes known electricity thieves as negative samples, summarizes the regular characteristics of the negative samples on the feature index items, and then evaluates the degree of difference between the detected users and the negative samples, thereby identifying abnormal users.

The core of the data-driven electricity theft detection method lies in two aspects: the selection of characteristic index items and the algorithm design. Electricity thieves usually have performance in specific aspects of metering data and user information. Commonly used characteristic index items in electricity anomaly detection mainly include the following categories:

(1) User payment records and credit information: users with bad credit have a significantly higher probability of abnormal electricity usage than general users. Credit evaluations such as the payment records of marketing systems and the on-site inspections times are commonly used characteristic index items [3–5].
(2) Consumer electricity consumption, volatility and fluctuation range: Electricity theft is often manifested as a trend of decline in electricity consumption. Most documents regard the load curve and fluctuation interval composed of daily or monthly electricity and their volatility as characteristic index items, and according to the abnormal changes in power consumption, the theft is identified [6–8].
(3) Reporting capacity utilization and power factor: In addition to abnormal power consumption, users may also leave traces in other aspects. For example, industrial and commercial users need to pay capacity electricity charges based on reported capacity, so there may be a certain proportional relationship between actual load and reported capacity [9]. In addition, power theft mainly focuses on reducing the power metering, and sudden abnormalities in the power factor may occur. Such indirect factors can also be used as characteristic index items.

On the basis of the aforementioned characteristic index items, researchers have explored the use of various algorithms to accurately identify abnormal electricity consumption. After completing the cluster analysis, outlier detection is required to identify outliers from the clustering results. The local outlier factor proposed in [10] can characterize the isolation degree of a sample by comparing the local density of an object and its neighbors in the data set, and can be used to detect outliers. [11] proposed an outlier identification method based on gridding local outlier factors, which significantly simplified the complexity of outlier analysis in low-density areas and improved the algorithm efficiency. Since there may be a strong correlation between the feature index

items, this paper also uses principal component analysis to reduce the dimension of the feature index items to improve the accuracy of the algorithm. Literature [7] uses Gaussian kernel function to improve the local outlier factor algorithm for user clustering on the basis of feature index dimensionality reduction, which better solves the problem of threshold setting for outlier judgment. In Irish commercial and residential user data. The test samples constructed above have a higher recognition accuracy. Literature [12] extended the horizontal load clustering between users to the vertical clustering of the users themselves, using the bulldozer algorithm for horizontal and vertical clustering, and using the user's electricity consumption behavior pattern. Aiming at the problem of high false alarm rate due to insufficient negative samples for electricity theft detection methods, literature [13] combined the ROC curve to optimize the classification threshold, and used the real-valued deep confidence network to detect the abnormal electricity consumption data for daily electricity data.

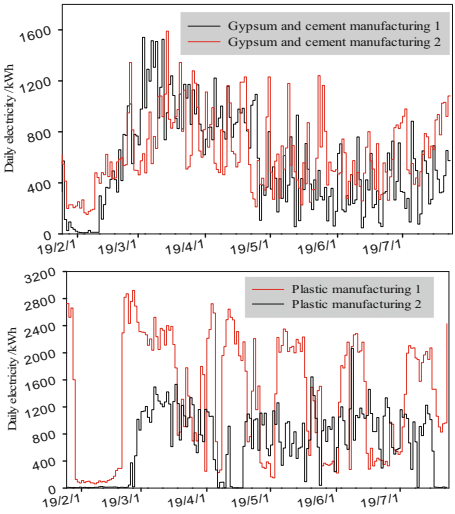



**Fig. 1.** Village tobacco roasting electricity consumption standardization curve.

It should be pointed out that the existing researches mostly design characteristic index items around abnormal changes in power consumption, which are prone to false alarms themselves for the following reasons. Existing power theft algorithms implicitly require users to have a basically stable power consumption, but from the Fig. 1, we can see that the power demand of a considerable number of users depends on the order demand, and large or trending fluctuations in power are normal. Industrial users with relatively stable electricity demand may have abnormal low electricity under external interference such as environmental protection inspection and safety inspection.

The aforementioned characteristic index items are distilled out of the industry background. Although this can better take into account the universality of users in different industries, the index items designed around power changes have the defects of insufficient reliability and sensitivity, which may easily lead to false alarms, and the defects of

this index item itself are difficult to solve from the algorithm design level. To the best knowledge of the authors, it is possible to extract industrial characteristic index items and detect abnormal electricity consumption by taking advantage of the similar characteristics of users in the same industry in terms of composition, consumption behavior and production scheduling mode of electric equipment.

## 2 Industry Characteristics Analysis of Flue-Cured Tobacco Users

Electric flue-cured tobacco is a typical case of promoting electric energy substitution strategy in rural areas recent years [13]. As electric roasting can automatically control the temperature of the roasting barn with high precision, it can not only reduce labor and fuel costs significantly, but also improve the quality of tobacco leaves. At present, electric flue-cured tobacco has been widely promoted and applied. In some tobacco producing areas, there are thousands of flue-cured tobacco in one county. These special properties are scattered in the fields, so it is difficult to conduct electricity inspection in a timely and effective manner.

Combined with the characteristics of the electricity consumption behavior of the flue-cured tobacco users, it is possible to set the characteristic quantity that can accurately describe the electricity consumption behavior in a targeted manner. In order to manifest the electricity consumption behavior of users clearly, the electricity consumption data of 4 typical flue-cured tobacco special properties in a certain place from June 1, 2016 to April 30, 2017 are selected below, and the annual daily electricity consumption data after standardization is plotted as shown in Fig. 2.

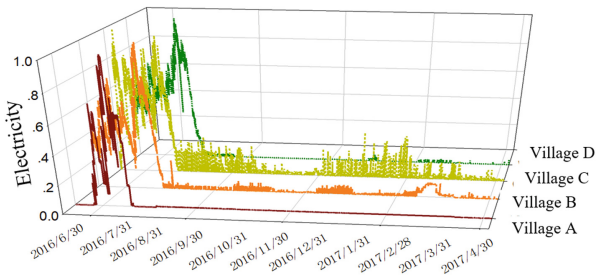



**Fig. 2.** Standardization daily electricity consumption in a year.

The characteristics of daily electricity consumption are analyzed as follows:

(1) The capacity of the flue-cured tobacco special properties is generally between 90~150 kW, and several barns are connected underneath, and each barn has a single-phase load, and the electric heating power is about 3~5 kW. Each barn can operate independently and does not have the characteristics of three-phase load balance. Therefore, it is not appropriate to use three-phase balance as a characteristic index to describe the electricity consumption behavior of flue-cured tobacco users.

(2) The flue-cured tobacco special properties mainly supplies the flue-cured tobacco load. The electricity consumption in the mature season of tobacco leaves in June and July reaches its peak, which is significantly higher than other periods. During the flue-cured tobacco season, the daily electricity consumption has obvious daily changes. From the Fig. 2 we can see that, taking the average daily power consumption of 4 special properties in June and July as the reference value, the calculated maximum daily power difference before and after the flue-cured tobacco season can reach 1.40, 0.36, 0.76, and 0.29, respectively. Because the flue-cured tobacco special properties load has obvious inter-day volatility during the flue-cured tobacco season, it is difficult to identify the abnormal electricity consumption based on the electricity abnormality.
(3) Other periods outside the flue-cured tobacco season are obviously light-loaded, with low electricity consumption and a stable trend, showing a continuous low electricity consumption state.

From the daily electricity data of the flue-cured tobacco special properties, it is difficult to identify abnormal electricity consumption based on the three-phase symmetry of the load or the inter-day electricity changes. It is necessary to combine the flue-cured tobacco technological process to extract characteristic index items from other time scales. The electricity consumption curve of the flue-cured tobacco special properties during one week of the flue-cured tobacco season is drawn as shown in Fig. 3. For the convenience of comparison, the weekly electricity curve of the commuter change in the local rural area is also drawn.

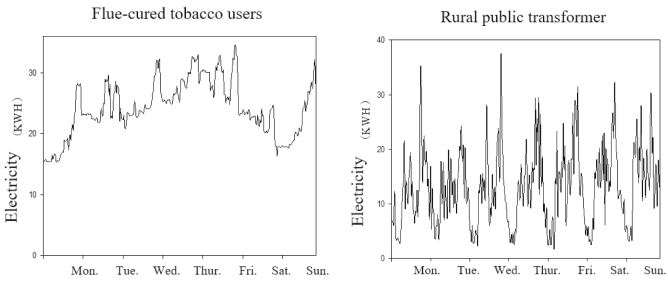



**Fig. 3.** Weekly power consumption of different types of users

It can be seen from the figure that the daily load curve of the public properties in rural areas has clear peaks and valleys according to the time changes of daily life. The electricity consumption at night is significantly lower than that during the day, which has outstanding daily cycle characteristics. The load of the common transformers in the station area fluctuates obviously on an hourly scale, and appears as a large number of burrs on the weekly electricity curve.

Although the daily power of flue-cured tobacco has obvious fluctuations, the load of flue-cured tobacco has strong continuity on an hourly scale. According to the technical requirements of tobacco leaf roasting, it takes about 7 days to roast tobacco leaves, during

which interruption of roasting will cause the tobacco leaves to re-moisture and economic loss. In order to keep the curing barn continuously dry and constant temperature, the flue-cured tobacco load remains basically stable. Only when the barn completes the roasting and moves out the cured tobacco leaves and moves into the to-be-cured tobacco leaves, there will be obvious load fluctuations. Therefore, the flue-cured tobacco users have relatively stable load on the hourly scale of the flue-cured tobacco season, and there will be no burr-like fluctuations and no load interruptions. Taking the two users in Fig. 4 as an example, taking a week's average daily power as a reference value, the flue-cured tobacco users' cumulative power fluctuation rate per week is 0.8202, while the Taiwan communal variable fluctuation rate can reach 7.2089, which is significantly higher than the former. Therefore, the cumulative fluctuation rate can better characterize the strong continuity of flue-cured tobacco load.

The load of flue-cured tobacco is limited by the technological process, with strong continuity, no daily periodicity, and the peak-valley difference between day and night is significantly smaller than that of ordinary production and living users.

## 3 First Section Construction of Industry Characteristic Index Items

### 3.1 Valley Peak Load Ratio

Please note that the first paragraph of a section or subsection is not indented. The first paragraphs that follows a table, figure, equation etc. does not have an indent, either. In order to adapt to the differences in load peak and valley periods in different regions and different seasons, the peak and valley periods can be calculated according to daily load data. When measuring at 30-min intervals, a day contains 48 points of measurement data. First, calculate the electricity consumption for 6 consecutive hours every 30 min. After sorting in ascending order, take the 6 h with the largest electricity consumption as the peak period and the 6 h with the smallest electricity consumption as the valley period. The power consumption during peak and valley periods and the peak-to-peak load ratio are calculated according to formulas (1), (2) and (3).

$$valley = \sum_{i=1}^{12} a_i^{valley} \tag{1}$$

$$peak = \sum_{i=1}^{12} a_i^{peak} \tag{2}$$

$$ratio = valley \Big/ peak \tag{3}$$

The peak load ratio of flue-cured tobacco season can describe the behavior characteristics of flue-cured tobacco users effectively. The trough-peak load ratios of the four typical flue-cured tobacco users in Fig. 2 during the flue-cured tobacco season are calculated daily, and their box-line diagrams are drawn as shown in Fig. 4.

It can be seen from Fig. 4 that the median peak-to-peak load ratios of the four typical flue-cured tobacco users are all between 0.8 and 1, and the range between the upper and lower quartiles is small, indicating that the peak-to-valley load difference during

the maturity period of tobacco leaves is relatively small. There are obvious differences among users in common residential areas, and the low-peak load ratio can effectively identify the characteristics of flue-cured tobacco load behavior.

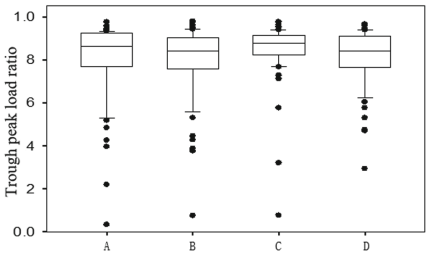



**Fig. 4.** Box line chart of electricity consumption ratio during peak and valley time span.

### 3.2 The Cumulative Volatility of Peaks and Valleys

Except for the trough-to-peak load ratio, the flue-cured tobacco load is relatively stable with a small cumulative change rate on an hourly scale, and can also be used to characterize the industry. Flue-cured tobacco users have a very low cumulative power fluctuation rate during the peak and low load periods, while other users may only show a lower cumulative power fluctuation rate during the night low-load period. To facilitate the multi-dimensional characterization of user behavior, the load can also be reduced. The cumulative volatility of power during peak and trough periods is collectively used as a feature item.

After calculating and determining the peak-valley period of each day according to the method described in the previous section, for 30-min interval measurement data, the cumulative value of peak and valley period fluctuations can be calculated according to formula (4) and formula (5).

$$cv_{peak} = \sum_{i=1}^{12} b_{i-1,i}^{peak} \tag{4}$$

Among them, $b_{i-1,i}^{peak} = |a_i - a_{i-1}|$

$$cv_{valley} = \sum_{i=1}^{12} b_{i-1,i}^{cvlley} \tag{5}$$

Among them, $b_{i-1,i}^{valley} = |a_i - a_{i-1}|$

In order to facilitate the comparison among users of different power consumption levels, after calculating the peak and valley load averages in formulas (6) and (7), according to formulas (8) and (9), the unified dimension of the peak and valley period fluctuation cumulative value is obtained after the value.

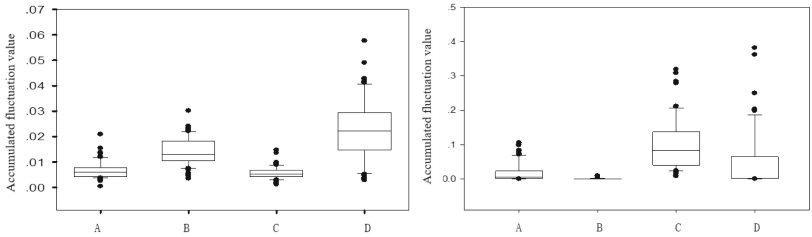
$$Mean_{peak} = \sum_{i=1}^{12} a_i \Big/ 12 \tag{6}$$

$$Mean_{valley} = \sum_{i=1}^{12} a_i / 12 \tag{7}$$

$$\overline{cv}_{peak} = cv_{peak} / Mean_{peak} \tag{8}$$

$$\overline{cv}_{valley} = cv_{valley} / Mean_{valley} \tag{9}$$

After analyzing the fluctuating accumulation values of four typical flue-cured tobacco special properties during the peak and valley periods of tobacco leaf maturity, the box plots are drawn as shown in Fig. 5a and b.




(a) peak load (b) valley time

**Fig. 5.** Accumulated variation of load curve.

The median cumulative volatility of the four typical users during the peak load period of the flue-cured tobacco season is between 0–0.02. The range between the upper and lower extensions of the boxes in Village B and Village D is slightly larger than that in Village A and Village C, and is stable within 0.03. There are outliers outside the tentacles of the box plot, but all samples including outliers are concentrated within 0.07. The median of the cumulative power volatility during the valley period is distributed between 0–0.1. All data points including anomalies are distributed within 0.4, and the overall distribution is relatively scattered.

## 4 Load Clustering Based on Nearest Neighbor Propagation Algorithm

### 4.1 A Subsection Sample Nearest Neighbor Propagation Algorithm

The nearest neighbor propagation algorithm is a clustering algorithm based on the transmission of nearest neighbor information. It has no need to specify the number of clusters, no need to select initial values. The cluster center point is the prominent advantage of the sample points that exist in the data set and the square and small errors of the clustering results, which are widely used in image, text and signal processing fields [14, 15].

The nearest neighbor propagation algorithm is based on the similarity matrix of the data set. At the initial stage, all samples are regarded as potential clustering center points, and the attraction is recursively transmitted along the node line until the optimal cluster

representative point set is found, so that the sum of similarities between all data points and corresponding cluster representative points is the largest. Among them, attractiveness is the degree to which data points are selected as cluster representative points of other data. The following describes the specific process of the nearest neighbor propagation algorithm in conjunction with Fig. 6.

a) The goal of clustering is to minimize the distance between the data points and the class representative points. Euclidean distance is selected as the similarity measure between data points. The similarity between any two points $x_i$ and $x_k$ is:

$$s(i,k) = -d_{i,k} = -\|x_i - x_k\| \tag{10}$$

b) Initialize the attraction matrix R and the attribution matrix A, use the attraction matrix R and the attribution matrix A to represent the two types of information between the data points, $r(i,k)$ represents the attraction from the point $x_i$ to the candidate representative point $x_k$, and indicates the degree to which $x_k$ is suitable as the class representative point of $x_i$; $a(i,k)$ is the degree of belonging from the point $x_k$ to $x_i$, which indicates the suitability of $x_i$ to select $x_k$ as the class representative point.

c) Calculate $r(i, j)$and $a(i, j)$

$$r(i,k) = \begin{cases} s(i,k) - \max\limits_{j \neq k}\{a(i,j) + s(i,j)\}, i \neq k \\ s(i,k) - \max\limits_{j \neq k}\{s(i,j)\}, i = k \end{cases} \tag{11}$$

$$a(i,k) = \begin{cases} \min\left\{0, r(k,k) + \sum_{j \neq k} \max\{r(j,k), 0\}\right\}, i \neq k \\ \sum_{j \neq k} \max\{r(j,k), 0\}, i = k \end{cases} \tag{12}$$

d) Iteratively update $r(i, j)$ and $a(i, j)$ to introduce the attenuation coefficient λ; the updated value is (1−λ) times the calculated value in this round plus λ times the previous round value.

$$r_{t+1}(i,k) = \lambda * r_t(i,k) + (1-\lambda) * r_t(i,k) \tag{13}$$

$$a_{t+1}(i,k) = \lambda * a_t(i,k) + (1-\lambda) * a_t(i,k) \tag{14}$$

e) Sum the attribution and attractiveness of each data point to determine the cluster center. Among them, for data point $i$, when $a(i,k) + r(i,k)$ achieves the maximum value, if $i = k$, the data point $i$ is determined to be the cluster center, if $i \neq k$, the data point is determined k is the cluster center.

$$k = \arg\max\{a(i,k) + r(i,k)\} \tag{15}$$

f) When the number of iterations of the cluster center is equal to the preset maximum number of iterations of the cluster center, clustering ends when the cluster center still does not change or the number of iterations equals the preset maximum number of iterations.

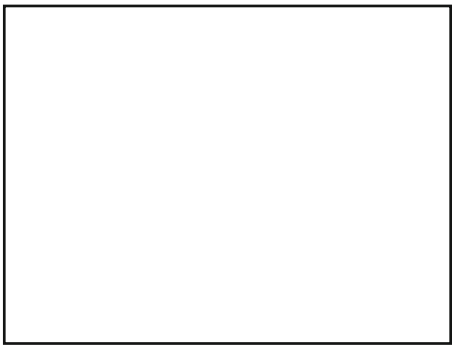


**Fig. 6.** Iterative diagram of AP clustering algorithm

### 4.2 Load Clustering Based on Nearest Neighbor Propagation Algorithm

The nearest neighbor propagation algorithm is used for cluster analysis of flue-cured tobacco users. First, obtain the electricity consumption data of each user in the same time period at equal intervals, and calculate the peak-to-peak electricity consumption ratio, the cumulative fluctuation rate of the peak-hour electricity consumption, and the cumulative fluctuation rate of the valley-hour electricity consumption. If the acquired electricity consumption data contains s days, each user will correspond to a 1 × (3 s) vector.

The similarity matrix is the initialization matrix of the algorithm. There is an $n \times n$ similarity matrix for $n$ users, and each value represents the similarity between two users. The diagonal element is recorded as preference, which is the similarity between the user and itself. The calculation result is 0 according to the Euclidean distance formula, so the preference value is the median of all similarity elements. The larger the value, the clustering that can be obtained the greater the number.

When updating attractiveness and attribution, attractiveness means that user $k$ is more suitable as the clustering center of $i$ than other users. For other users $k'$, $s(i, k')$ represents the similarity between users $i$ and $k'$. Redefine the index $a(i, k')$ to indicate the degree of recognition of user $i$ to $k'$ as its cluster center. By calculating the value of $s(i, k') + a(i, k')$, $k'$ can be obtained as the suitability of the cluster center of $i$; among all other users $k'$, find $\max\left\{a(i, k') + s(i, k')\right\}$, and finally calculate $s(i, k) - \max\left\{a(i, k') + s(i, k')\right\}$ get user $k$'s attractiveness to $i$, where $k \neq k'$; For attribution, first calculate user $k$'s attractiveness to other users $r(i', k)$, and then the cumulative summation represents the attractiveness of user $k$ to other users $\sum_{i' \neq k} \max\left\{r(i', k), 0\right\}$, plus $r(k, k)$, this value reflects how much user $k$ should not be divided into other cluster centers, and $\min\left\{0, r(k, k) + \sum_{i' \neq k} \max\left\{r(i', k), 0\right\}\right\}$ can be get.

Finally, sum the user attraction and attribution to find the cluster center that maximizes $a(i, k) + r(i, k)$. If $i = k$, then the cluster center is i itself, if i $\neq$ k$i \neq k$, the cluster center is $k$.

In order to evaluate the results of the clustering algorithm, the Purity index is used as the evaluation standard, which is defined as the proportion of the correct clustered data to the total data.

$$Purity(W, X) = 1 \Big/ N \sum \max_j \left|\omega_k \cap x_j\right| \tag{16}$$

Among them, $W = \{w_1, w_2, \cdots, w_k\}$ is the set of clusters, $w_k$ represents the k-th cluster set, $X = \{x_1, x_2, \cdots, x_n\}$ is the data set, and $x_j$ represents the j-th data, N represents the total number of data. The value of this index is between 0–1. When it is 0, the clustering result is completely wrong, and when it is 1, the clustering result is completely correct. The Purity index can directly reflect the performance of the clustering algorithm, and the accuracy of the constructed feature index can be judged according to the evaluation index to describe the electricity consumption behavior of flue-cured tobacco users.

## 5 Numerical Simulation

From the electricity consumption information collection system, 200 users whose industry attribute is the tobacco industry are selected to conduct electricity abnormality detection. In the actual system, there may be errors in industry attributes caused by registration errors and user business changes. In order to simultaneously test the clustering algorithm's ability to detect users with incorrect industry attributes, it also involved 20 rural public changes. In order to fully characterize the characteristics and trends of the electricity consumption of flue-cured tobacco users in the mature season of tobacco leaves, the electricity consumption data is selected from the 30-min interval data of the 2016 flue-cured tobacco season (from June 1 to July 30).

The parameters of the nearest neighbor propagation clustering algorithm are set as follows: the attenuation coefficient λ is 0.5, the maximum number of iterations is 500 times, the maximum number of iterations that the cluster center does not change is 50 times, and the reference degree is set to all values in the similarity matrix The median. In addition, K-means, FCM fuzzy C-means clustering algorithm, and DBSCAN are also used to cluster the data set. Because users may be special properties in normal flue-cured tobacco and abnormal flue-cured tobacco, and public changes in rural life, the number of clusters is set to 3 in the K-means and FCM algorithms. Each clustering algorithm uses Purity index to evaluate the performance of the algorithm, and the evaluation index can be shown in Table 1.

The nearest neighbor propagation clustering algorithm divides the test user data into 5 categories. Except for flue-cured tobacco users and rural living areas, it is found that 4 of the 20 rural areas that are mixed are actually brick factory users and are individually identified as one. In addition, the two flue-cured tobacco special properties are separately identified as two categories; the number of clusters is more than the expected 3 categories. The Purity index of cluster recognition is 0.9818, which has the highest correct rate among the four clustering algorithms. Among them, 4 flue-cured tobacco users are classified as rural living areas. In-depth understanding shows that the local capacity is insufficient due to the lack of distribution and the residents' load is

connected to the flue-cured tobacco transformation, which leads to clustering errors. It needs to be pointed out that the neighbor propagation algorithm parameters setting is simple, and the clustering effect is stable under different parameter settings.

The nearest neighbor propagation algorithm can not only accurately cluster and distinguish normal/abnormal flue-cured tobacco users based on user characteristic indicators, but also identify non-cured tobacco users with other industry attributes mixed with test data.

The parameter setting has a significant impact on the clustering effect of the DBSCAN algorithm. After a large number of parameter setting tests, users can also be divided into 5 categories under optimal conditions, and the Purity index can reach up to 0.9545.

K-means and FCM clustering algorithms are clustered according to the preset number of clusters of 3 types, and the clustering effect is obviously inferior to DBSCAN and neighbor propagation clustering algorithms.

**Table 1.** Results of clustering algorithm evaluation indicators.

| Algorithm | Number of clusters | Purity Metrics |
|---|---|---|
| K-means | 3 | 0.7955 |
| FCM | 3 | 0.8455 |
| DBSCAN | 5 | 0.9545 |
| AP | 5 | 0.9818 |

**Table 2.** Cluster contain users and division accuracy.

| | User | Accuracy |
|---|---|---|
| Cluster 1 | Flue-cured Tobacco Specialized User (194 households) | 100% |
| Cluster 2 | Rural public change (16 households), Flue-cured Tobacco Specialized User (4 households) | 80% |
| Cluster 3 | Brick factory (4 households) | 100% |
| Cluster 4 | Abnormal Flue-cured Tobacco Special Change User A (1 household) | 100% |
| Cluster 5 | Abnormal Flue-cured Tobacco Special Change User B (1 household) | 100% |

According to the classification of neighbor propagation clustering, the 4 flue-cured tobacco special properties were mistakenly classified as the rural public properties. The specific conditions are listed in Table 2. In order to compare the characteristic indicators of different clusters, the 60-day average values of the peak cumulative volatility, trough cumulative volatility and trough peak load ratio of various cluster centers are listed in Table 3, and each cluster center is listed on June 1st. The electricity consumption curve up to July 30 is drawn as shown in Fig. 7. Each sub-graph corresponds to the set of flue-cured tobacco users, the set of communal transformers in rural areas, the set of brick factory users, the abnormal flue-cured tobacco user A, and the abnormal flue-cured tobacco user B. It can be seen from the chart that:

(1) The characteristic items of various cluster centers are significantly different. Cluster center 1 is a flue-cured tobacco user with strong load continuity and low cumulative volatility, and the cumulative volatility is significantly lower than other cluster centers; center 2 is a rural station with obvious daily cycles. Center 3 is a brick factory user with strong cumulative volatility, and the cumulative volatility is significantly higher than that of all other cluster centers; center 4 and 5 have obvious index differences, and they are two abnormal flue-cured tobacco users. From the curve characteristics and index items, the clustering results are reasonable.
(2) Cluster 4 and cluster 5 are applied to the load of abnormal flue-cured tobacco users A and B. Among them, abnormal user A only has normal power consumption during the first few days of the flue-cured tobacco season in June, and abnormal user B has continuous electricity consumption for 5 days separated in mid-June, which violates the rules of continuity requirements for flue-cured tobacco, which is significantly different from typical flue-cured tobacco users, and can be judged as abnormal electricity consumption.

**Table 3.** 30-day mean of cluster center features.

| | Peak cumulative volatility | Trough cumulative volatility | Trough peak load ratio |
|---|---|---|---|
| Cluster 1 | 0.922 | 0.827 | 0.696 |
| Cluster 2 | 5.185 | 5.975 | 0.610 |
| Cluster 3 | 13.006 | 14.014 | 0.434 |
| Cluster 4 | 6.142 | 0.517 | 0.001 |
| Cluster 5 | 4.767 | 3.761 | 0.334 |

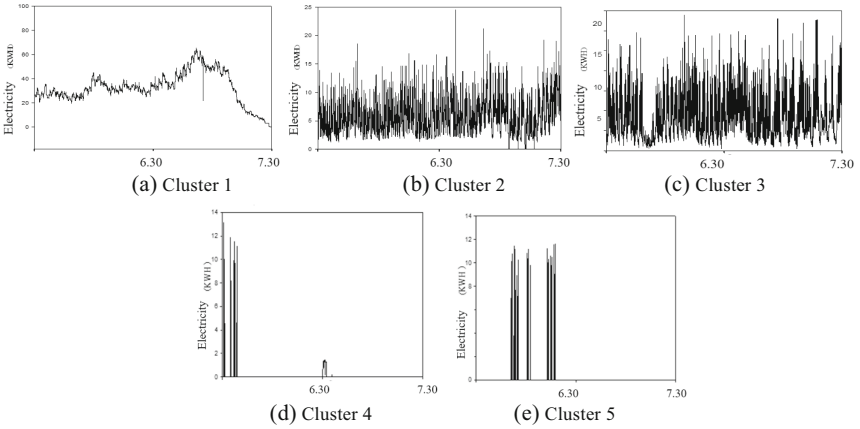



**Fig. 7.** Load profile of Clustering center

## 6 Conclusion

In this paper, users in the same industry have similar characteristics of electricity consumption behavior, and propose an abnormal electricity detection method considering the electricity consumption characteristics of users in the industry. Combined with users of flue-cured tobacco, explore the extraction of characteristic index items by sub-sector and detect abnormal electricity consumption. The characteristic index items that can accurately describe the electricity consumption characteristics of flue-cured tobacco users are adopted. The nearest neighbor propagation algorithm is used to perform cluster analysis on the actual flue-cured tobacco special change and the rural public transformer according to the extracted characteristic index items. The analysis results show that the proposed method not only accurately identifies the flue-cured tobacco special change users with abnormal electricity consumption, but also can effectively distinguish them. Specialized users of non-flue-cured tobacco mixed in due to wrong industry attributes.

## References

1. Hao, R., Ai, Q., Xiao, F.: Architecture based on multivariate big data platform for analyzing electricity consumption behavior. Electr. Power Autom. Equip. **37**(8), 20–27 (2017)
2. Cheng, C., Zhang, H., Jing, Z., et al.: Study on the anti-electricity stealing based on outlier algorithm and the electricity information acquisition system. Power Syst. Protect. Control **43**(17), 69–74 (2015)
3. Su, S., Li, K., Yan, Y., et al.: The resident load consumption pattern classification model based on density spatial clustering and gravitational search algorithms. Electr. Power Autom. Equip. **38**(1), 129–136 (2018)
4. Nagi, J., Yap, K.S., Tiong, S.K., et al.: Nontechnical loss detection for metered customers in power utility using support vector machines. IEEE Trans. Power Deliv. **25**(2), 1162–1171 (2010)
5. Chen, Q., Zheng, K., Kang, C., et al.: Detection method of abnormal electricity consumption: review and Prospect. Autom. Electr. Power Syst. **42**(17), 189–198 (2018)

6. Sun, Y., Li, S., Cui, C., et al.: Detection method of power data outliers based on Gaussian kernel function. Grid Technol. **42**(5), 1595–1604 (2018)
7. Marcelo, Z., Edgard, J., Marcelo, P., et al.: A tunable fraud detection system for advanced metering infrastructure using short lived-patterns. IEEE Trans. Smart Grid **10**(1), 830 (2019)
8. Monedero, I., Biscarri, F., León, C., et al.: Detection of frauds and other non-technical losses in a power utility using Pearson Coefficient, Bayesian networks and decision trees. J. Electr. Power Energy Syst. **34**, 90–98 (2012)
9. Breuning, M.M., Kriegel, H.P., Ng, R.T., et al.: LOF: identifying density-based local outliers. ACM Sigmod Int. Conf. Manage. Data **9**(2), 93–104 (2000)
10. Zhuang, C., Zhang, B., Hu, J., et al.: Anomaly detection for power consumption patterns based on unsupervised learning. Proc. CSEE **36**(2), 379–387 (2016)
11. Feng, Z., Tang, W., Wu, Q., et al.: Users' consumption behavior clustering method considering longitudinal randomness of load. Power Autom. Equip. **38**(9), 39–44 (2018)
12. Zhang, C., Xiao, X., Zheng, Z.: Electricity theft detection for customers in power utility based on real-valued deep belief network. Power Syst. Technol. **43**(3), 1083–1091 (2019)
13. Peng, X., Lai, J., Chen, Y.: Application of clustering analysis in typical power consumption profile analysis. Power Syst. Protect. Control **42**(19), 68–73 (2014)
14. Li, L., Liu, T.: PV power forecasting based on AP-ESN. Electr. Power Autom. Equip. **36**(7), 41–46 (2016)
15. Hang, W., Jiang, Y., Liu, J., et al.: Transfer affinity propagation clustering algorithm. J. Softw. **27**(11), 2796–2813 (2016)