Anomaly Recognition and Correction Algorithm for Big Data of Distribution Network Load

Tao Huang¹, *, Ziqiang Wang¹, Wei Wang¹, Qiang Zhang¹ and Yanwei Chen²

¹China Southern Power Grid Power Dispatching Control Center, Guangzhou, China
²Beijing Tsingsoft Technology Co., Ltd, Beijing, China

*Corresponding author: huangtao2@csg.cn

Abstract. With the continuous improvement of the intelligent degree of distribution network, load information presents the growth of big data level, large-scale, mixed, inaccurate monitoring or collection of load data often appear, which brings some difficulties to the scheduling and forecasting work. Therefore, it is necessary to identify and correct these abnormal data. Based on the analysis of the causes and distribution characteristics of different types of abnormal load data, this paper identifies and modifies the abnormal data based on the improved density estimation algorithm, so as to realize the rapid processing and repair of the distribution network load big data. Finally, the power load data of a province in 2020 is selected to clean the load data. The results show that the algorithm proposed in this paper can quickly and accurately realize the identification and correction of abnormal load data.

1. Introduction

With the improvement of China's power network intelligence level and the deepening of power market reform, power load data has gradually become one of the most widely used data, which plays an irreplaceable role in power system operation, power consumption analysis and power demand control. However, the power load data will deviate from its normal value due to various reasons, so it is particularly important to identify and correct the abnormal values in the load data before applying the power load data. In this paper, these deviation data are collectively referred to as the abnormal value of power load.

With the increasing popularity of smart meters in China, massive power load data are collected and uploaded to the centralized control center. In the face of such a huge amount of data, it is unrealistic to identify outliers manually, and it must be automatically completed by using efficient and reliable data processing methods. Reference [1] studies the quality of power load data recording in smart grid environment, focusing on the causes of abnormal load values and the current detection methods. In the aspect of load outlier’s identification, reference [2] uses the sample clustering depth to identify the distorted load shape, and uses the traditional t-test method to test the abnormal data points. In reference [3], the problem of insufficient pattern drift of neural network is improved, and a model of abnormal data recognition and adjustment is proposed. In reference [4], the idea of fuzzy soft clustering is integrated into the neural network model, and combined with the radial basis function (RBF) network, the abnormal data recognition and cleaning model is formed. In reference [5], a
dynamic data processing technology based on multi-source information is proposed to correct load outliers by using the characteristics of energy management system with multiple data sources.

So far, in the aspect of identification and correction of abnormal load data, most of the research work only considers the horizontal or vertical continuity of load, that is, processing in one dimension, which has certain limitations. Therefore, this paper considers the horizontal and vertical continuity of load at the same time. Specifically, firstly, the load data is arranged into two-dimensional data by date. Then, the density estimation method is used to identify the abnormal data as a whole. Finally, the identified abnormal data is corrected. The actual power load data of Guangdong Province is also used for calculation. The simulation results show that this method is effective.

2. Characteristic analysis of abnormal load data

In the actual operation of power system, the complexity and unknowability of random factors lead to the generation of abnormal data with certain randomness, and the types of load data are complex and no single. Generally speaking, according to the different causes of abnormal load data, abnormal data can be roughly divided into the following categories [6]:

2.1. missing data

The missing data in the abnormal load data of power system is generally caused by some line maintenance blackouts, load shedding blackouts or temporary faults of SCADA system. It is mainly manifested as the random single vacancy or continuous vacancy value in the database unit table, which is shown in the load curve as shown in Figure 1.

![Figure 1 load curve of missing data](image)

Figure 1 shows the first abnormal type, in which the black solid line represents the actual collected load curve, and the blue dotted line represents the ideal load curve. It can be seen from the figure that the actual load curve is missing at points 5, 13, 14, 15 and 16, which destroys the continuity of the whole curve.

2.2. Distorted data

This type of data usually refers to the data with a certain deviation from the actual value in a certain range. Due to the influence of internal or external random factors of the system, the load data fluctuates randomly in a small range, resulting in the distortion of the load data. There are two types of distortion: glitch load and impact load. The main cause of glitch load is the small zigzag fluctuation of data amplitude caused by the internal noise of the system. The impact load data is mainly caused by unexpected events, such as the obvious decrease or increase of power load caused by extreme weather. Figure 2 shows the abnormal curve of common deformities.
The abnormal data marked at points 73-83 in the figure are impact load data, which shows that the load value in this period has a sharp drop due to the influence of some factors; the abnormal data marked at points 139-147 are glitch load data, and the load data in this period has a zigzag distortion.

2.3. Mutation data
The abrupt abnormal load data mainly refers to the maximum and minimum value phenomenon in the load curve, which generally shows that the load data of the power system is lower than the valley value of the curve or higher than the peak value of the curve at some time points. This kind of abnormal data type is similar to the distorted data. The distinguishing point is the magnitude of load data mutation, which is usually based on the peak value and valley value of the load data on that day. Therefore, the mutation data belongs to the burr data with excessive mutation. The load curve is shown in Figure 3.

Obviously, the maximum and minimum values of the load curve appear at point 14 and 36 respectively. The maximum value of the load is 1.1105 MW and the minimum value is 0.9213 MW. The peak value and valley value of the load curve are 1.1 MW and 0.9328 MW respectively. It can be seen that the maximum value is significantly higher than the peak value of the whole load curve and the minimum value is significantly lower than the valley value of the whole load curve.
3. Abnormal data recognition method based on density estimation

3.1. Description of data density estimation method
(1) Suppose there is a two-dimensional data set Z with the total number of data points M.

(2) To generate a data set S called seed group, the number of seeds N should be determined in advance, and the distance between each seed and its adjacent seeds should be the same. In addition, the range of seed group should be able to include the data set Z.

(3) Each data point \( z_j \) \((j \in \{1, 2, ..., M\})\) is attached with a seed adsorption counter \( c_i \) with an initial value of 0, which is used to accumulate the number of seeds adsorbed by the data point.

(4) For each seed \( s_i \) \((i \in \{1, 2, ..., N\})\), the distance between it and each data point of the data set Z is calculated respectively, assuming that the nearest data point to the seed \( s_i \) is \( z_k \). The process of using Euclidean distance to determine the order of the nearest data point \( z_k \) to the seed \( s_i \) (i.e., subscript \( k \)) can be expressed as follows:

\[
k = \arg \min \{ ||s_i - z_j||^2 \}
\]

Where \( i \in \{1, 2, ..., N\}, j \in \{1, 2, ..., M\} \), \( \arg \) is the abbreviation of argument. The meaning of formula (1) is to take the value of \( j \) that minimizes the value of objective function as \( k \).

(5) According to equation (1), determine the data point \( z_k \) closest to the seed \( s_i \), and add 1 to the seed adsorption counter \( c_k \) attached to the data point. If there are \( p \) data points which are equal to and closest to the seed \( s_i \), they are allocated to these data points in equal proportion, that is, the seed adsorption counter of each data point closest to the seed \( s_i \) is accumulated by \( 1/p \).

(6) For each seed in the seed group S, the nearest data point is determined according to formula (1), and then the value of the seed adsorption counter at the corresponding data point is updated according to the above rules until all the seeds are calculated.

3.2. Recognition principle
Each data point is accompanied by a seed adsorption counter to accumulate the number of seeds adsorbed by each data point. Thus, there are two conclusions:

(1) If the value of the seed sorption counter of a data point is large, it indicates that the data point adsorbs more seeds, that is, there are not many data points in its neighborhood that compete with it to share these seeds, and the data point density is low;

(2) If there are many data points in the neighborhood of a data point, there will be fierce competition between the data point and its surrounding data points when adsorbing seeds, each data point will adsorb fewer seeds.

A lower density of data points indicates a lower probability of data points appearing in their neighborhoods, which can classify data points with a seed sorption counter value higher than a set value as bad data.

3.3. Specific calculation process
First, the calculation method of the data center is given.

At time \( t \), the density center model is as follows:

\[
\hat{m}(t) = \frac{1}{N} \sum_{i=1}^{N} \omega_i y_i
\]

Where \( y_i \) is the observed value of the load at time \( t \), and \( \omega_i \) is the weight of the point \((t, y_i)\).

In nonparametric regression analysis, the \( \omega_i \) pass formula (3) gives:
In the formula: $\text{Kern}_h(l)$ denotes the kernel density function with scale parameter $h$; $l$ is marked as the central load position and $l=N+1$; $l_i$ is the load observation position marker and $l_i \in [1, N]$.

The specific kernel functions selected in this paper are as follows:

$$Kern_h(l) = \frac{1}{h} Kern\left(\frac{l}{h}\right)$$

(4)

Generally, for a given $F$-distribution data, outlier identification is to confirm that some observations are distributed in the outlier data domain. For any confidence level $\alpha$, the anomalous data field for the distribution $F$ of parameter $\theta$ at the $\alpha$ level is defined as [7]:

$$out(\alpha, \theta) = \left\{ x : x < Q_{\alpha/2}(\theta) \text{ or } x > Q_{1-\alpha/2}(\theta) \right\}$$

(6)

$Q_{\alpha/2}(\theta)$ is the $\alpha/2$-quantile of the distribution function $F(\theta)$. At time $t$, the relationship between observed load data and central load is as follows:

$$y_i = \hat{m}_h(t) + \epsilon_i$$

(7)

Where $\epsilon_i$ is the deviation between the central load and the load observations. It is assumed that the deviation term $\epsilon_i$ is independently identically distributed and follows a standard normal distribution with a mean of 0 and a variance $\sigma^2$. Because the modified sample variance is an unbiased estimate of the population variance, the estimate of variance $\sigma^2$ is:

$$\hat{\sigma}^2 = \frac{1}{N-1} \sum_{i=1}^{N} (y_i - \hat{m}_h(t))^2$$

(8)

The exception data field is:

$$out(\alpha, \hat{\sigma}) = \left\{ y : |y - \hat{m}_h(t)| > Q_{1-\alpha/2} \hat{\sigma} \right\}$$

(9)

Where $Q_{1-\alpha/2}$ is the $100 \times (1 - \alpha/2)$ quantile of the standard normal distribution.

After obtaining the anomaly data domain, follow the steps below to identify the load anomaly values.

Step 1: For the basic power mode dataset, according to the load level of a subset of daily load data, select the corresponding abnormal data domain to identify the load outliers.

Step 2: For special power mode datasets, load level matching is required first. For the "peak-valley-peak" power mode, according to the proportion of time periods of high and low loads, the eigenvectors...
formed by high or low loads are selected to match the load level, and the abnormal data fields of corresponding load levels are used to identify the load outliers. If the match fails, go to step 3.

Step 3: For daily loads in special power mode datasets that have not yet been identified for load outliers, the maximum upper and minimum lower bounds of all outlier data domains are used to identify load outliers for load data.

4. Abnormal data correction method

In terms of load abnormal value correction, the conventional weighted average method and the weighted average method based on load level mapping are used for different situations [8]. When the days to be corrected can find the days with similar load levels, the conventional weighted average method is used. The conventional weighted average method is to select the load data of the daily load adjacent to the day to be corrected at the same time as the reference. In this paper, the load data of similar days with similar load levels at the same time are selected as reference, and the weighted average value is used as the correction value. The correction formula is as follows:

\[
L_{d,t} = \lambda_1 y_{d-1,t} + \lambda_2 y_{d-2,t} + \ldots + \lambda_m y_{d-m,t} \\
\sum_{g=1}^{m} \lambda_g = 1
\]

(10)

In the formula: \(m_a\) is the number of similar days selected; \(y_{d-g,t}\) and \(\lambda_g\) are the load observation value at time \(t\) on the \(d-g\) day and the weight of the influence to the correction value. When no days with similar load levels are found on the days to be corrected, the weighted average method that introduces the load level mapping relationship is adopted, and the whole process is corrected in chronological order. The correction formula is as follows:

\[
L_{d,t} = \lambda_1 f(y_{d-1,t}) + \lambda_2 f(y_{d-2,t}) + \ldots + \lambda_m f(y_{d-m,t}) \\
\sum_{g=1}^{m} \lambda_g = 1
\]

(11)

In the formula: \(m_b\) is the number of adjacent days selected; \(f(y_{d,g,t})\) and \(\lambda_g\) are the mapped load value at time \(t\) on day \(d-g\) and the influence weight of the correction value to be treated, respectively.

5. Case study

Taking the annual load of Guangdong Province in 2020 as a sample, some data error points are artificially set, and the abnormal data rate is 1.25%. Traditional neural network algorithm is used as a comparison.

Fig. 4 and Fig. 5 show the corrected values of point 5, 6, 7, 8 on January 15 and point 4 and 19 on March 10 using this method (Method A) and traditional neural network (Method B), respectively. The black curve in the graph represents the actual load data corresponding to January 15 and March 10. The blue curve is the artificially modified abnormal load data. The red curve and the green curve represent the corrected values of the abnormal load point based on method A and method B, respectively. By comparing the change trend between the red and green curves and the original load curve, it is found that the load curve corrected by method A is closer to the original normal load value in value. Obviously, whether it is to correct the continuous multipoint abnormal load data or the load data of a single abnormal point, the correction accuracy of the abnormal load data correction method proposed in this paper is much higher than that of using the neural network to correct the abnormal load data.
Figure 4 Continuous outlier correction results

Figure 5 Single point outlier correction results

Table 1 shows the error comparison results for the correction of abnormal load data in 2020. The average relative errors of method A for correction of continuous point and single point outliers are 3.58% and 1.36%, respectively, while that of method B is 5.14% and 2.80%. Compared with the correction errors of method B for continuous points and single point outliers, the correction errors of method A proposed in this paper for continuous points and single point outliers are reduced by 1.56% and 1.44%, respectively. Therefore, the accuracy of outlier correction method based on density estimation algorithm is much higher than that based on traditional RBF.

| Correction Method | Average Relative Error % |
|-------------------|--------------------------|
|                   | Continuous error | Single point error |
| Method A          | 5.14             | 2.80              |
| Method B          | 3.58             | 1.36              |

6. Conclusions
The method of identifying and correcting abnormal data based on density estimation can effectively identify and correct continuously abrupt or continuously missing data points, and the recognition process is based on the original data as a whole, avoiding the shortcomings of the existing horizontal comparison method. The verification of the proposed method with actual data of Guangdong power system shows that the method in this paper is feasible and effective.

Acknowledgments
This work was financially supported by Key science and technology projects of China Southern Power Grid Corporation ZDKJXM20190020. Fund Project: key science and technology project of China Southern Power Grid Corporation, Project No.: zdkjxm20190020
References

[1] CHEN Wen, ZHOU Kaile, YANG Shanlin, et al. Data quality of electricity consumption data in a smart grid environment [J]. Renewable and Sustainable Energy Reviews, 2017, 75: 98-105.

[2] Alex R, Alessavdro L. Machine learning. Clustering by fast search and find of density peaks [J]. Science, 2014, 344(6191): 1492-1496.

[3] GU Min, GE Liangquan, QIN Jian. Identification and justification of dirty electric load data based on modified ART2 network [J]. Automation of Electric Power Systems, 2007, 31(16): 70-74.

[4] ZHANG Xiaoxing, CHENG Jiyun, ZHOU Quan, et al. Dynamic intelligent cleaning for dirty electric load data based on data mining [J]. Automation of Electric Power Systems, 2005, 29(8): 60-64.

[5] YE Feng, HE Hua, GU Quan, et al. Bad data identification and correction for load forecasting in energy management system [J]. Automation of Electric Power Systems, 2006, 30(15): 85-88.

[6] Huang S J, Lin J M. Enhancement of Anomalous Data Mining in Power System Predicting-Aided State Estimation [J]. IEEE Transactions on Power Systems, 2004, 19(1): 610-619.

[7] Manitsas E, Singh R, Pal B C, et al. Distribution System State Estimation Using an Artificial Neural Network Approach for Pseudo Measurement Modeling [J]. IEEE Transactions on Power Systems, 2012, 27(4): 1888-1896.

[8] Fatemi S B, Mobasheril M R, Abkar A A. Improving the Accuracy of Multispectral Image Clustering by Means of a New Initializing Method [J]. Journal of the Indian Society of Remote Sensing, 2016, 44(4): 643-650.