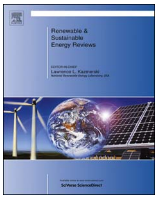
Renewable and Sustainable Energy Reviews xx (xxxx) xxxx–xxxx

Contents lists available at ScienceDirect

Renewable and Sustainable Energy Reviews

journal homepage: www.elsevier.com/locate/rser



Data quality of electricity consumption data in a smart grid environment

Wen Chen[a], Kaile Zhou[a,b,*], Shanlin Yang[a,b], Cheng Wu[c]

[a] School of Management, Hefei University of Technology, Hefei 230009, China
[b] Key Laboratory of Process Optimization and Intelligent Decision-making of Ministry of Education, Hefei University of Technology, Hefei 230009, China
[c] Department of Industrial Engineering, Tsinghua University, Beijing 100084, China

ARTICLE INFO

Keywords:
Electricity consumption data
Data quality
Outlier detection
Outlier data
Smart grid

ABSTRACT

With the increasing penetration of traditional and emerging information technologies in the electric power industry, together with the rapid development of electricity market reform, the electric power industry has accumulated a large amount of data. Data quality issues have become increasingly prominent, which affect the accuracy and effectiveness of electricity data mining and energy big data analytics. It is also closely related to the safety and reliability of the power system operation and management based on data-driven decision support. In this paper, we study the data quality of electricity consumption data in a smart grid environment. First, we analyze the significance of data quality. Also, the definition and classification of data quality issues are explained. Then we analyze the data quality of electricity consumption data and introduce the characteristics of electricity consumption data in a smart grid environment. The data quality issues of electricity consumption data are divided into three types, namely noise data, incomplete data and outlier data. We make a detailed discussion on these three types of data quality issues. In view of that outlier data is one of the most prominent issues in electricity consumption data, so we mainly focus on the outlier detection of electricity consumption data. This paper introduces the causes of electricity consumption outlier data and illustrates the significance of the electricity consumption outlier data from the negative and positive aspects respectively. Finally, the focus of this paper is to provide a review on the detection methods of electricity consumption outlier data. The methods are mainly divided into two categories, namely the data mining-based and the state estimation-based methods.

1. Introduction

At present, with the rapid and stable development of the economy, the demand for energy is constantly expanding. As an important form of energy, electricity plays an important role in the development of modern economy and society [1,2]. With the deepening reform of energy sector and the continuous progress of energy technologies, electric power will play an increasingly indispensable role. Wind energy, solar energy and other new energy sources can be converted into electricity. The electric power industry is one of the most important basic energy industries in the development of national economy. With the development of economy and society, electricity demand continues to expand. It promotes the expansion of electricity consumption market and stimulates the development of electric power industry. In recent years, information technology for electric power industry has also been developed. The construction of smart grid and the application of emerging IT technology make the scale of electric power big data resources continues to grow [3,4]. Construction and application of the smart grid have stimulated the accumulation of electric power grid operation data, production management data and electricity consumption data [5,6]. These accumulated data contain redundant, missing and outlier data, resulting in serious issues of electric power data quality.

Electric power data quality problems are not just faced by China. Many countries in the world have these problems. At present, many countries in the world are in the reform of electricity market and vigorously promoting the construction and application of smart grid. The scale of electricity consumption data collected by advanced metering infrastructure (AMI) in smart grid is becoming increasingly huge in many countries. The quality of these electricity big data has a direct impact on the accuracy and effectiveness of electric power system management and application based on data analysis, which further affects the safety and reliability of the whole power system [7,8]. Therefore, data quality of electricity consumption data is an important research and application issue for the development of power industry in many countries.

The quality of electricity consumption data is a core issue in the process of electricity data mining, and it plays an important role in

[*] Corresponding author at: School of Management, Hefei University of Technology, Hefei 230009, China.
E-mail addresses: zhoukaile@hfut.edu.cn, kailezhou@gmail.com (K. Zhou).

http://dx.doi.org/10.1016/j.rser.2016.10.054
Received 18 April 2016; Received in revised form 3 September 2016; Accepted 30 October 2016
Available online xxxx

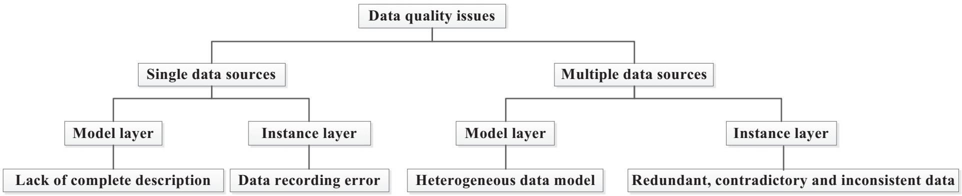


Fig. 1. Classifications of data quality issues [28].

energy big data analytics [9,10]. At present, some researches on data mining focus on the data mining algorithm and ignore the data quality processing before data analysis. Some mature algorithms have requirements on the data set, such as data integrity and data redundancy [8,11–13]. However, in real life, electricity consumption data are incomplete, redundant and ambiguous, which are inconsistent with the requirements of many data mining algorithms [14]. In addition, noise data seriously affect the efficiency of data mining algorithms and result in invalid induction [15–17]. The improvement of electricity consumption data quality has become a key issue in the realization of electric power data mining system.

This paper mainly discusses the data quality issues of electricity consumption. The structure of this paper is as follows. Section 2 introduces the meanings and classifications of data quality. Section 3 describes the data quality of electricity consumption, including the characteristics and classifications of electricity consumption data quality. We focus on the description of three kinds of electricity consumption data quality issues, i.e., noisy data, incomplete data and outlier data. Then in Section 4, we analyze the reasons and significances of electricity consumption outlier data. The detection of electricity consumption outlier data are introduced in Section 5, which can be divided into two categories, namely the data mining-based and the state estimation-based electricity consumption outlier data detection methods. Finally, Section 6 summarizes the full text.

2. Data quality

With the rapid development of Information and Communication Technologies (ICTs), large amounts of data have been accumulated from all walks of life. High data quality is the basic condition of data analysis and mining. In order to discover knowledge from data to support decision-makings, we must ensure the data quality. Therefore, it is necessary to understand the meanings of data quality deeply.

2.1. Data quality management

There are many definitions of data quality. Saha [18] pointed out that data quality refers to the recognition of outlier data and elimination of error data before data is loaded into data center. Huang et al. [19] believed that data quality is the degree of data suitable for use. Roger and Mangiameli [20] focused on the four properties of data quality, namely accuracy, integrity, consistency and timeliness. They pointed out that data quality is the consumer of information. Alizamini et al. [21] believed that data quality is a complex non-structural concept and data refinement process. As can be seen from the above, there is no one uniform consensus on the definitions of data quality. From different point of view, the definitions of data quality are also different. In this paper, we define data quality management as the process of removing the noise, identifying the outlier and processing the incomplete from raw data. Eventually, data quality is higher, and the results of data analysis are more accurate.

Data quality is a comprehensive concept of multidimensional factors, which can be described and evaluated from six aspects, i.e., accuracy, integrity, consistency, self-consistency, availability and timeliness [22]. The data accuracy is the core of data quality, which refers that data must correctly and truly reflect the actual business. Data business descriptions are also reasonable and accurate. The data integrity means that there are no field and record deletion in the recorded data. The recorded data can fully describe the recorded business. The data consistency means that data is logically consistent [23]. It mainly includes three sub-indexes, namely concept, range and format of consistency [24]. The data self-consistency is that data should meet the constraints, which describes the relationships between data [25,26]. Data must satisfy the relationships to each other and cannot be contradictory. The data availability refers to the available degree of data. That is, data should be easy to access, understand and use [27]. The data timeliness implies that data can play a role in the required time. Outdated data has a negative impact on the data quality and indirectly affects the results of data analysis.

2.2. Classifications of data quality issues

Aebi [28] proposed that data quality issues can be divided into four categories, namely the single data source model layer, single data source instance layer, multiple data source model layer and multiple data source instance layer, as shown in Fig. 1.

The single data source issues can be analyzed on two aspects, i.e., the model and instance layer. From the perspective of model layer, lacking complete descriptions and low-grade model designs are main issues. Although database has a complete description of data and model design, it may also have some data quality issues, such as lacking unique or referential constraints. The main issue of instance layer is human errors, such as spelling errors, duplicate records and so on [29–31].

The multiple data source issues are more complex. The multiple data source model layer has heterogeneous data model issues, such as name and structure conflicts [32]. For the multiple data source instance layer, it has redundant, contradictory and inconsistent data.

3. Data quality of electricity consumption data

3.1. Electricity consumption data

At present, large amounts of electricity consumption data have been accumulated. It is difficult for people to directly discover hidden knowledge behind these data which are heterogeneous and inconsistent [33]. In order to discover knowledge, how to ensure data quality is the most important step. To improve the quality of electricity consumption data, its characteristics should be fully considered. With the development of information technology and smart grid, electricity consumption data present the following characteristics.

(1) The volumes of data have increased rapidly. With the development of smart grid, many electric power companies have established Big Data and cloud computing centers to process the data. So it makes the available data increased.
(2) Data transmission and processing speed have greatly accelerated due to the establishment of intelligent equipment and system [34].
(3) Management is more precise. The data mining technology makes the analysis and management of electric energy meters data more precise.

The volumes of electricity consumption data are huge. Also, the types of electricity consumption data are various and the value is high. Through analysis of these data can dig out the high value knowledge of electric power enterprises to enhance its level of profitability and control. According to the experts' analysis, when data utilization rate has increased by 10%, electric power enterprises can increase the profit of 20–49% [35]. Hence, we must improve the utilization rate of electricity consumption data. However, if data are redundant and confused, it will increase the difficulty of obtaining effective knowledge. If data are wrong and outdated, it will result in incorrect data analysis results. Therefore, ensuring high quality of electricity consumption data is the basis of data applications.

The electricity consumption data quality issues can be divided into three categories, namely noise data, incomplete data, and outlier data. Noise data refer that data is against the business logic [36–39]. Thus, it can affect the results of data analysis. Incomplete data refer to the data of missing attribute values in the data source, or the data that deviated from statistical characteristics [40,41]. This type of data is generated by special events in the case of system operating normally. Therefore, it is likely to contain a lot of useful knowledge [42]. Moreover, identification and management of outlier users are conducive to carry out personalized marketing services for electric power companies. In a word, how to identify the noise, incomplete and outlier data in electricity consumption data is an important process to improve the data quality.

3.2. Noise data in electricity consumption data

Noise data refer to the data that is difficult to be understood and translated by machines, e.g., unstructured text [43], which results in low data quality. Noise data in electricity consumption data include logical errors and inconsistent data [44–46].

(1) The logical errors mean that actual data is not consistent with the attribute data, which violates the business rules or logic [47,48]. For instance, the monthly electricity consumption data of a household includes 32 days. This is obviously beyond the maximum number of days in the month. So it is clearly illogical.
(2) Data in the following situations are called inconsistent data. First, data do not conform to a particular format [49], e.g., different date formats. Second, data deviate from the distribution of a column, e.g., negative electricity consumption records. Third, when comparing the different attributes of same records, data lack significances [50,51]. For example, the date of user's electricity consumption is earlier than the date of house purchased by the user.

3.3. Incomplete data in electricity consumption data

Due to the complexity of smart grid environment, there are some certain degrees of incomplete in electricity consumption data. The existing incomplete data processing methods can be divided into three categories, namely delete tuple, data filling and non-processing [52,53]. Delete tuple method refers to the deletion of incomplete data records. Thus, we can obtain a data set with no incomplete data. This method has an advantage of simple operation. But it is only applicable to data sets that included a small amount of incomplete data. Data filling method refers that we fill incomplete data through the appropriate algorithms based on no missing data. At present, researchers have put forward many incomplete data filling methods [47,53,54], e.g., artificial filling method, average value filling method, special value filling method and regression method [54]. However, they are not applicable to incomplete data in electricity consumption because these incomplete data may contain some useful knowledge. If we fill it up, the purpose of outlier detection will lose [54,55].

3.4. Outlier data in electricity consumption data

Outlier data refer that data deviate from most data in the data set. There are two reasons for the generation of outlier data, namely objective and subjective factors. Some conditions changed or unknown factors emerged are objective factors [56,57]. Subjective factors refer to human factors. Therefore, evaluation of data quality should consider the background of outlier data [56–58].

Outlier data in electricity consumption data may be removed as noise data. But it may contain important knowledge of data units. Hence, the detection of outlier data is very important in research of data quality. For outlier data, there are many detection methods, which are based on statistical model, distance and deviation. Lueebber and Grimmer [59] put forward a method of using data audit to detect outlier data automatically, which was called data quality mining (DQM).

In fact, outlier data in electricity consumption data are not all error data. So we should make further analysis combined with relevant knowledge according to the actual situations [60]. These data may contain some important knowledge which is ignored. Therefore, it is very meaningful to study the electricity consumption outlier data. The following paper analyzes the causes and significances of electricity consumption outlier data. Afterwards, the detection methods of electricity consumption outlier data are summarized, namely based on data mining and based on state estimation methods.

4. Generation of electricity consumption outlier data

Outlier data have an important influence on accuracy, completeness and self-consistency of data quality. At the same time, it has important events information of electric power grid, such as power rationing, equipment failures and so on [61]. Thus, it is of great significance to identify, analyze and deal with outlier data.

4.1. Causes of electricity consumption outlier data

In residential or industrial electricity, it will produce large amounts of consumption data. Most consumption data are normal. However, there are also some outlier data. The generation of outlier data usually has the following reasons.

(1) When electric power system is in operation, measurement and transmission processes of data acquisition system may generate outlier data. For example, outlier changes of data in the transmission are caused by the failure of data transmission system [62]. Thus, data may lose or confuse in the transmission process.
(2) The data acquisition system is normal. But outlier changes of electric load are caused by special events, such as: line outage maintenance [63]. The occurrence of special events usually makes the electricity consumption data in a certain period of time as a null value. This is a kind of outlier phenomenon.
(3) Significant electricity reforms make electricity consumption habits changed, such as: Ladder-type price [64]. Ladder-type price sets a base to residents, and more than the base price will increase the electricity price [65]. This will lead to changes of electricity habits and result in outlier electricity consumption.
(4) These conditions may also produce outlier data, such as: records errors, artificial forgery and so on. User's electricity consumption data may hide the user's electricity behavior habits. For data mining of electricity consumption data, on the one hand, it can understand the users' demands for personalized service; on the other hand, it can detect the users' stealing behavior and protect the benefits of enterprises.

4.2. *Significances of electricity consumption outlier data*

Complex electric power system contains large amounts of real-time data. Data is accurate or not, which determines the safety and reliability of electric power system. Outlier data may affect the normal operation of electric power system and even threaten the security of entire system. Hence, in order to ensure the stable and safe operation of electric power system, it has important significances to extract and detect these outlier data from the original data. The significances of electricity consumption outlier data have two aspects on negative and positive influences.

The negative influences include the following aspects. Firstly, the presence of electricity consumption outlier data will reduce the accuracy of assessment. Then, the results of data mining cannot accurately reflect the characteristics of data. Finally, outlier data affect the judgment and decision of electric power system dispatcher [65,66]. It even threatens the safe operation of system. If outlier data cannot be correctly identified and effectively corrected, they will provide false prediction as a reference [67]. This affects the accuracy and reliability of prediction results. Furthermore, if outlier data are used as modeling data, it will interfere with the changing rules and influence the training accuracy. If outlier data are used as a predictor of test results, it can lead to erroneous judgement [68].

Outlier detection of electricity consumption data is to find out the relatively sparse and isolated outlier data patterns which are hidden in massive data [69]. In the early stage of data set preprocessing, we usually put outlier data as noise data. Although outlier detection finds the hidden data in the data set as the main purpose, outlier data mining is more valuable and meaningful than other types of mining [70]. Because one hundred thousand normal records are likely to cover only one rule, but ten outlier records are likely to cover ten different rules [68–71]. Outlier detection of electricity consumption data may provide us more important information, so that we can find some real and unexpected knowledge, which can help us to understand the consumer behavior, capture the theft, find system vulnerabilities and failures and improve service quality [67,72].

5. Detection of electricity consumption outlier data

The experts put forward different research methods for outlier detection in the electric power system, which can be divided into two categories, namely based on data mining and based on state estimation methods.

5.1. *Outlier detection methods based on data mining*

Data mining obtains the previously unknown knowledge from massive, incomplete, noisy, fuzzy and random data. The data mining process can be divided into three steps, including data preparation, i.e., data integration, data selection, data preprocessing and data conversion, data mining and interpretation evaluation [70–72]. Outlier detection method based on data mining can be divided into four methods, which are based on neural network, based on fuzzy theory and cluster analysis, based on the Gap Statistic Algorithm and based on time series analysis methods.

5.1.1. *Methods based on neural network*

Neural network imitates the structure and working mechanism of human brain to build a computational model [73]. Due to its good robustness, self-organization, self-adaptive, parallel processing and distributed storage characteristics, neural network is very suitable for solving the problems of classifications model in data mining [74]. The method based on neural network is shown in Fig. 2. It is composed of three layers, namely input layer, hidden layer and output layer. The input layer defines the value of all input attributes for data mining. The hidden layer is the position of assigning weights to various input probabilities. The output layer represents the predictable attribute value of data mining model [73,74].

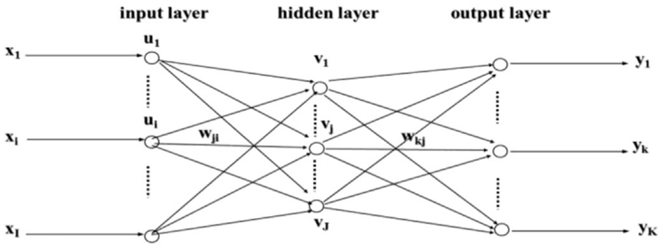


Fig. 2. Diagram of neural network [79].

Rakhshani et al. [74] proposed a method of detecting and identifying the outlier data in electric power system based on two level neural networks. The input of first level neural network was the original measurement data, and the output was the state of system. The second level neural network used the output of first level neural network as input, and the output was the predicted value of measured data. Two level neural networks were used to predict the difference between the predicted value and measured value. If the difference was big, it was outlier data. Teeuwsen [75] constructed a back propagation (BP) neural network. He used measured data as the training sample of the back propagation neural network for real-time monitoring. This method can correctly detect and identify outlier data to realize fault diagnosis of electric power system. Souza et al. [76] constructed a neural network based on GMDH (Group Method of Data Handling). They used regular information as input variables to detect and identify outlier data of electric power system. In order to correctly identify outlier data in real time monitoring, the reverse propagation neural network was introduced. But the detection effect was restricted by training samples, and it was difficult to achieve the desired goal. Zhang et al. [77] proposed to establish the inverse propagation neural network. They used it to eliminate the interference before the estimation. Then the purpose of identifying the various forms of power grid measurement errors was achieved.

The neural network continuously studies and adjusts the characteristics pattern of subject by training, which constructs a characteristic profile with adaptive characteristics. This is the key of outlier detection to form a characteristic profile of a user or system. But using neural network to identify electricity consumption outlier data also has some shortcomings. First, the selected data will be directly related to the final results. Hence, data quality of selected data is an important influence factor. Second, the identification process of this method requires the selection of appropriate thresholds. But the choice of thresholds is subjective. So the results are influenced by human factors to a certain extent. Third, this method easily causes residual pollution and residual submersion that result in missed detection and false detection.

5.1.2. *Methods based on fuzzy theory and cluster analysis*

Clustering collects the physical or abstract objects into similar objects composed of multiple classes. Firstly, data are divided into several classes. Then, the classes are a collection of data objects. Eventually, these objects are similar to the objects in the same cluster and different from the objects in other clusters [78–80]. The most commonly used algorithm in clustering analysis is k-means algorithm based on partition. According to the number of K, the algorithm randomly selects K initial cluster centers, and it does not stop iteration. When the minimum value of objective function is reached, we obtain the final results [81]. The objective function is usually defined as the square error criterion. The definition is as the Eq. (1):

$$E = \sum \frac{k}{i=1} \sum_{p\hat{I}c_i} |p - m_i|^2 \tag{1}$$

E represents the sum of mean square deviation of all objects. *p* is a

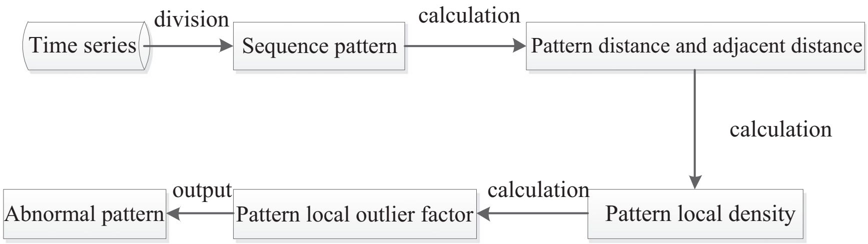


Fig. 3. Outlier detection model based on time series [48,92–94].

point in the object space. The mean value of clustering c_i is expressed by m_i [82].

Gao and Shan [83] proposed a method based on the ISODATA (Iterative self-organizing data analysis algorithm), which could accurately identify the electric power system outlier data and avoid the influence of initial clustering center on classification results. Sun et al. [69,84] introduced the concept of fuzzy mathematics into the detection and identification of outlier data in electric power system. On the basis of studying on fuzzy clustering theory, the detection and identification of outlier data were carried out by several fuzzy clustering models. According to the analysis and comparison of its results, the outlier data identification system based on the fuzzy clustering theory was developed. Huang and Lin [85] found the cluster centers of normal and outlier measurement data by using the fuzzy clustering analysis. This method realized the cleaning of outlier data of power load. Feng et al. [86,87] applied outlier data mining to electric power load forecasting. They put forward by combining the advantages of hierarchical clustering method and information entropy principle to select the basic parameters of clustering process. Also, they used artificial neural network to extract the characteristic curve of load curve for correcting outlier data.

The existing clustering algorithms can be divided into two categories, namely hard clustering and fuzzy clustering algorithms. In hard clustering algorithm, the membership degree of a class can only be 0 or 1. It is easy to produce local extreme value. For the membership degree of a certain category, the sample value of fuzzy clustering is in the interval of [0,1]. The sum of membership degree for all categories samples is 1. Comparing with hard clustering algorithm, the fuzzy clustering algorithm is more in line with the objective reality. But the speed of fuzzy clustering is slower than hard clustering. The outlier detection of fuzzy clustering theory can avoid the residual pollution and residual submersion. However, the shortcoming of this method is to determine the membership degree subjectively.

5.1.3. Methods based on the Gap Statistic Algorithm (GSA)

In 2000, Tibshirini et al. [88] presented the gap statistics data mining algorithm (Gap Statistic Algorithm, GSA). Firstly, we preprocess the data by neural network. Then, the optimal number of clusters is determined automatically by the gap statistic algorithm. Finally, clustering analysis is carried out on the normal data and outlier data to realize the detection and identification of outlier data. The core of GSA is to compare the natural logarithm of clustering discrete degree with a reference value [88]. Then, we determine the optimal number of clusters [88]. An equation is defined here:

$$gap(k) = E\ \ln[W_{refj}(k)] - \ln[W(k)] \tag{2}$$

In Eq. (2), E is the mathematical expectation of reference data. With the change of K, when $gap(k)$ is maximum value, K is considered to be the most appropriate number of clusters [88].

Wu et al. [89] provided a new method of the elbow judgment based on the GSA algorithm. This method combined with GSA was used to identify outlier data in electric power system. It had high accuracy and reliability, and computing speed also improved greatly. Yang et al. [90,91] proposed that the GSA clustering algorithm mixed on the advantages of neural network was applied to the identification of outlier data in electric power system. Compared with the traditional state estimation methods, it can avoid the residual pollution and residual submersion.

When sample data are detected, the traditional GSA must consider the errors. If sample data are large, the computation of simulation errors is large. The improved GSA is still based on the guiding ideology of the traditional GSA. But the "Gap" statistic of the traditional algorithm is modified. The root causes of errors are eliminated fundamentally. Therefore, "Gap" can reflect the facts truly and objectively. At the same time, it is unnecessary to estimate and calculate the standard deviation and simulation errors. It reduces the computational complexity and improves the operation speed.

5.1.4. Methods based on time series analysis

As shown in Fig. 3, it is a detection method of outlier data based on time series analysis. The principle of this method is to identify the time series of each state by using time series model. After that, the outlier patterns of data are detected. In addition, judging outlier data is the "useful data" of extracting fault information or the "useless data" of being cleaned [48].

Yan et al. [58] proposed a method of outlier detection based on time series analysis and unsupervised learning, which achieved outlier detection from data evolution process and data association. Through time series model, Messina and Vittal [92] could detect the dynamic trend and frequency of electric power network in time. But the deficiency was that it could not be used in the detection of a large number of data streams. Guo et al. [93] presented to detect outlier data in load data by time series modeling. Sheng et al. [94] tested the noise points and missing values in the state data transmission equipment by using iterative test method. This method could correct the outlier data in the iterative process.

When outlier data is generated by outlier state of equipment, the time series model is used to extract the effective fault information, which can effectively reflect dynamic change of original time series and adapt to the characteristics of power equipment state data. The time series analysis method has a limitation. That is, the series value has the time attribute and strict order. Outlier detection for disordered data is not mature.

In addition to the four mainstream methods mentioned above, there are some researchers leading the genetic algorithm into electric power industry. With the improvement of the complexity of electric power system, genetic algorithm plays an important role in solving the problem of electric power system with its unique characteristics and advantages. Genetic algorithm is widely used in power grid planning, fault recovery and so on [95]. Genetic algorithm is a search algorithm based on genetic principle. Through gene selection, reproduction, crossover and variation, the algorithm is optimized in order to approach the optimal solution [95,96].

When we apply fuzzy set in outlier detection, the determination of membership function depends on the expert's domain knowledge, which is quite subjective. Therefore, Su et al. [96] used genetic algorithm to optimize the parameters in order to search the best membership function. This outlier detection method was verified by data of electric power system. Victor et al. [97] applied genetic algorithm to intrusion detection systems, which could improve the speed and efficiency of detection. This method could be applied to the intrusion detection of electric power system. Xu et al. [98] introduced genetic algorithm into the field of electric power system planning. Abdulwhab et al. [99] used genetic algorithm for fault location in power system. According to the characteristics of distribution network, fault location was performed, and the global optimal solution was found.

Using genetic algorithm in outlier detection, the advantage is that the initial group of parameter sets can be determined randomly, which reduces the dependence on the experts' domain knowledge. However, for massive data, the application of genetic algorithm is a complex and time-consuming process.

5.2. Outlier detection methods based on state estimation

The data of electric power system may deviate from true values due to the fault of systems and human beings. This leads to the emergence of outlier data, which affects the subsequent state estimation. The state estimation is to improve the data accuracy by using the redundancy of real-time measurement system and to eliminate error information caused by the random disturbance [100–102]. Thus, the state of system can be estimated and predicted. The outlier detection method based on state estimation includes residual searching method, estimation identification method and non-two criteria method [102–104]. These methods are based on the assumption that the weighted residual or standard residual of eigenvalues obeys a probability distribution. According to a certain confidence level, a threshold value is determined to test the hypothesis. Then, the outlier data are found. The found outlier data can be eliminated or reduced the weight of measurement data. In the end, the new state estimation is got [105,106].

The basic idea of residual searching method is to queue residual (weighted residual or standard residual) from large to small. On the basis of this, the maximum value of residual is removed, and the state estimation is renewed. If there are suspicious data in the residual detection process, the above processes are to be repeated [107]. If the test is successful, the residual search identification process should be successful. However, it takes much time to perform multiple state-estimation calculation. So it cannot be used in large scale electric power system identification of outlier data [108,109].

Table 1 is a comparison of the residual search method and estimation identification method. From the functional point of view, the residual search method is aimed at single outlier data, and the estimation identification method is based on multiple outlier data. From the perspective of calculation speed, the estimation identification method is faster than the residual search method. For the complexity of procedure, both are very simple. For the correction of state estimation, the residual search method is to renew the state estimation, and the estimation identification method is to modify directly. The storage cost of residual search method is low. Relatively speaking, the storage cost of estimation identification method is high because it needs to store the W matrix.

Table 1
Comparison between residual search method and estimation identification method [105–110].

Index	Method	
	Residual search	Estimation identification
Function	Single outlier data	Multiple outlier data
Calculation speed	Slow	Fast
Complexity of procedure	Simple	
Correction of state estimation	Do state estimation again	Correction directly
Cost of storage	Low	Need to store W matrix

Non-two criteria method is a very typical method for the detection and identification of state estimation, which is used to calculate the residual error by using the non-two criteria estimator in the estimation calculation process. According to the measurement residual, adjusting the corresponding weight achieves the purpose of detecting outlier data [110,111].

Ye et al. [60,112] studied the reasons of outlier data in EMS (Energy Management System) of electric power system dispatch center. In this paper, outlier data were processed by using dynamic multiple source method with total added value. Zhang et al. [113] presented a method to solve absolute measurement errors of the demarcation point voltage. Huang et al. [114,115] proposed a method for identification of distributed outlier data, which was realized by using distributed residual method. Jabr [116] proposed to improve the convergence of state estimation by using least square method. Finally, the identification of outlier data was performed by using the residual error.

Outlier detection method based on state estimation assumes that measurement errors follow normal distribution. Then, the outlier measurement data are identified by the residuals based on hypothesis test. The detection and identification of state estimation calculation can detect multiple outlier data at the same time, and the efficiency is high. But outlier data detection method based on state estimation has some deficiencies. In the identification process, the repeated estimation needs more time, and the speed is slow. So it is not conducive to the detection of a large amount of data. When there are massive outlier data in the measured data, it is easy to cause errors. Thus, the accuracy of test results is low. The biggest disadvantage of this method is that there will be residual pollution and residual submersion, which affects the identification effect.

6. Conclusion

Electric power system has accumulated a large amount of electricity consumption data, which inevitably have outlier, redundant and incomplete data, affecting the quality of electricity consumption data. In the process of electricity big data application and sharing, ensuring data quality plays a key role in power system decision-making. In this study, we focus on the data quality and outlier detection of electricity consumption data.

(1) This paper analyzes and summarizes the traditional outlier detection methods, including the residual search method, non-two criteria method, estimation identification method and so on. The disadvantage of these methods is the emergence of residual pollution and residual submersion, resulting in missed detection and false detection. Traditional outlier data identification algorithms always use the least square method, and multiple state estimations in the identification process are needed, which is usually time consuming. Therefore, there are some limitations in the practical application of power system.

(2) Compared with the traditional outlier detection methods, some new theories have been developed, such as fuzzy mathematics and data mining algorithms. It has been demonstrated that data mining based methods have higher predictive and better detection ability for potential problems. Moreover, this kind of methods can effectively avoid the residual pollution and residual submersion and improve the detection speed and accuracy. Hence, the application of advanced data mining technology has a good prospect in the detection and identification of outlier data in electric power system.

There are still many issues worthy of studying further in detection and identification of electricity consumption outlier data.

(1) The development of smart grid makes it more convenient to collect the electric power data in real time. Therefore, identifying and

detecting real-time updated data are the focus of our next research.

(2) Outlier data in smart grid may be generated by a variety of factors, such as electrical equipment failures, network parameter errors and network intrusion. The identification and localization of these outlier data remain to be further studied.

(3) The evaluation system of outlier data identification remains to be further studied. At present, there are still no perfect evaluation systems to evaluate the effect of outlier data identification.

Acknowledgement

The authors would like to thank the editor and the anonymous reviewers for their comments and suggestions, which have improved the quality of the paper significantly. This work is supported by the National Natural Science Foundation of China (71501056, 71601063), the Fundamental Research Funds for the Central Universities (JZ2016HGTB0728), Anhui Provincial Natural Science Foundation Program (1608085QG165), Anhui Provincial Philosophy and Social Science Planning Project (AHSKQ2015D42), and the Foundation for Innovative Research Groups of the National Natural Science Foundation of China (No. 71521001).

References

[1] Zhou K, Yang S, Shen C, et al. Energy conservation and emission reduction of China's electric power industry. Renew Sustain Energy Rev 2015;45:10–9.
[2] Zhou K, Yang S. Demand side management in China: the context of China's power industry reform. Renew Sustain Energy Rev 2015;47:954–65.
[3] Zhou K, Fu C, Yang S. Big data driven smart energy management: from Big data to Big insights. Renew Sustain Energy Rev 2016;56:215–25.
[4] Zhou K, Yang S, Shao Z. Energy internet: the business perspective. Appl Energy 2016;178:212–22.
[5] Han J, Xu L, Dong Y. Data quality review. Comput Sci 2008;2:5–12.
[6] Rye KS. A study on data quality management maturity model. Electric. Power Syst Res 2005;13, [920-15].
[7] Panahy PH, Sidi F, Affendey LS, Jabar MA. The impact of data quality dimensions on business process improvement. Int J Electr Power Energy Syst 2014;9:1773–82.
[8] Chandola V, Banerjee A, Kumar V. Anomaly detection: a survey. ACM Comput Surv 2009(3):15–58.
[9] Zhou K, Yang S, Shen C. A review of electric load classification in smart grid environment. Renew Sustain Energy Rev 2013;24:103–10.
[10] Zhou K, Yang S, Chen Z, et al. Optimal load distribution model of micro grid in the smart grid environment. Renew Sustain Energy Rev 2014;35:304–10.
[11] Chen b. Bad data identification power system research. Power Syst 2010.
[12] Li H, Li W. A new method of power system state estimation based on wide-area measurement system. Ind Electron Appl 2009;43:2065–9.
[13] Roberto B, Francesco B. Outliers detection in multivariate time series by independent component analysis. Neural Comput 2007;5:1903–8.
[14] Yuri AW, Shardt XuY, Steven X. Quantization and data quality: implications for system identification. J Process Control 2016:40.
[15] Lucas A. Corporate data quality management: from theory to practice. Ind Electron Appl 2010:85.
[16] Huang Y, Cheng F. Automatic data quality evaluation for the AVM system. IEEE Trans Power Appar Syst 2011;24:1–8.
[17] Munawar M, Salim N, Ibrahim R. Towards data quality into the data warehouse development. Int J Electr Power Energy Syst 2011:5.
[18] Saha B, Srivastava D. Data quality: the other face of big data. J Data Inf Qual 2014;22:30–1.
[19] Huang K, Lee Y, Wang R. Quality information and knowledge management. Advances in Neural Information Processing Systems, 38; 1998. p. 1672–83.
[20] Blake H, Mangiameli Roger M. The effects and interactions of data quality and problem complexity on classification. J Data Inf Qual 2011;18:20–1.
[21] Alizamini FG, Pedram MM, Alishahi M, Badie K. Data quality improvement using fuzzy association rules. Int J Electr Power Energy Syst 2010;15:3–8.
[22] Lawrence H, Cavuto D, Papavassiliou S. Adaptive and automated detection of service anomalies in transaction oriented WAN's: network analysis, algorithms, implementation and deployment. IEEE Trans Power Appar Syst 2000;67:1876–87.
[23] Xing H, Zhang D. Methods to improve the accuracy of fault time power acquisition system. China CSEE 2011;16:88–95.
[24] Zhang L, Han J, Zhao E. Chinese power enterprise adaptability of the electric power system reform management structure model. China Manag Sci 2008;2:165–71.
[25] Ban X, Ning S, Xu Z. Novel method for the evaluation of data quality based on fuzzy control. J Syst Eng Electron 2008;3:606–10.
[26] Zeng M, Wu J, Liu C, Wang T. Considering multiple factors assessment of the residents of the price ladder. East China Electr Power 2012;3:359–62.
[27] See J, Carr W, Collier SE. Real time distribution analysis for electric utilities. IEEE Trans Power Deliv 2008;4:61–5.
[28] Aebi D, Perrochon L. Towards Improving data quality. IEEE Trans Power Syst 1993;10.
[29] Almutiry O, Wills G, Alwabel A, Crowder R, WaIters R. Toward a framework for data quality in cloud-based health information system. Int J Electr Power Energy Syst 2013;13:258–68.
[30] Boufares F, Ben A. Heterogeneous data-integration and data quality: overview of conflicts. International. J Electr Power Energy Syst 2012:13.
[31] Shahriar MS, Anam S. Quality data for data mining and data mining for quality data: a constraint based approach in XML. IEEE Trans Inst Electr Eng Jpn 2008:56.
[32] Bastos MR, Martini JS, De Almeida JR, Viana S. Data integration: quality aspects. Adv Neural Inf Process Syst 2010:23.
[33] Moslehi K, Kumar R. A reliability perspective of the smart grid. IEEE Trans Smart Grid 2010(1):57–64.
[34] Madnick Stuart, Zhu H. Improving data quality through effective use of data semantics. Data Knowl Eng 2006;78:460–75.
[35] Zhang L, Wang X, Li X. Power system operation information acquisition and maintenance technology. Beijing: China Electric Power Press; 2010.
[36] Sequeira RE, Gubner JA. Blind intensity estimation from shot-noise data. IEEE Trans Power Appar Syst 1997;45.
[37] Hao S, Zhou X, Hong S. A new method for noise data detection based on DBSCAN and SVDD. Int J Electr Power Energy Syst 2015;39:1042–9.
[38] Chen C, Jin J. A soft computing technique for noise data with outliers. Electr Power Syst Res 2004;25:1956–62.
[39] Ren J, Tang T, Yuan Y. Bayesian networks parameter learning based on noise data smoothing in missing information. Adv Neural Inf Process Syst 2012:46.
[40] Wang H, Li M, Chen B. The research of outlier data cleaning based on accelerating method. Ind Electron Appl 2010;44:835–43.
[41] Zhi W, Yu L, Li F, Yang S. Case base maintenance based on outlier data mining. Ind Electron Appl 2005;75:765–77.
[42] Jun Z, Hong W. A new pretreatment approach of eliminating abnormal data in discrete time series. IEEE Trans Power Energy 2005:98.
[43] Amini A, Saboohi H, Teh Ying Wah. A multi density-based clustering algorithm for data stream with noise. IEEE Trans Power Appar Syst 2013;12:643–55.
[44] Sen B, Qin Z, Li X, Li Z. The application and research of noise data acquisition with wireless network. IEEE Trans Power Deliv 2009:87.
[45] Hu K, Li L, Lu Z. A cleaning method of noise data in RFID data streams. Int J Electr Power Energy Syst 2013;6:1044–8.
[46] Wang S, Li Y, Liu N, Wang S. Data cleaning algorithm of noise data in attribute level. Comput Eng 2005;9:86–7.
[47] Bi F, Geng Z, Li H. Multiple models fusion for pattern classification on noise data. Ind Electron Appl 2012;16:324–34.
[48] Cao J, Diao X, Chen S, Shao Y. Data cleaning and general system framework. Comput Sci 2012;3:207–11.
[49] Erhard R., Hong-Hai D. Data cleaning: problems and current approaches. IEEE Data Engineering Bulletin, 4; 2000. p. 3-13.
[50] Yong Y, Trouve A. A non-linear k-means algorithm and its application to unsupervised clustering. IEEE Comput Appl Power 2002;45:135–45.
[51] Morto A, Anokhin P, AcarA C. Utility-based resolution of data inconsistencies. In: Proceedings of the 2004 International workshop on information quality in information systems. New York: ACM, 76; 2004. p. 35–43
[52] Wang S, Lin S, Lu Y. Analysis of incomplete data problem with Bayesian network. J Tsinghua Univ 2000;9:65–8.
[53] Li J, Li P, Shu K. RMINE: a rough set based data mining prototype for the reasoning of incomplete data in condition-based fault diagnosis. J Intell Manuf 2006;1:163–76.
[54] Qin Y, Zhang S, Zhu X. Pop algorithm: kernel-based imputation to treat missing data in knowledge discovery from databases. Expert Syst Appl 2009;2:2794–804.
[55] Little R, Rubin D. Statistical analysis with missing data. John Wiley & Sons; 2002.
[56] Chen J, Dai T, Zhang N, Gan D, Wang K. The deterministic contract decomposition in abnormal load data identification and correction. Autom Electr Power Syst 2009;6:21–4.
[57] Yu F, Xu H, Wang Q, Li G. A new method of abnormal data detection on traffic flow of extra-long highway tunnel. Int J Electr Power Energy Syst 2010;17:1245–66.
[58] Yan Y, Sheng G, Chen Y, Jiang X, Zhi H, Du X. Based on the data analysis of power transmission and transformation equipment condition data anomaly detection method. Proc CSEE 2015;1:52–9.
[59] Lueebber D, Grimmer U. Systematic development of data mining based data quality tools. Power Deliv 2003;36:5643–54.
[60] Ye F, He H, Gu Q. Bad data identification and correction for load forecasting EMS. Autom Electr Power Syst 2006;15:85–8.
[61] Jun Z, Hong W. A new pretreatment approach of eliminating abnormal data in discrete time series. Int J Electr Power Energy Syst 2005;18:1982–99.
[62] Xu M, Lu Z, Qiao Y, Wang N, Zhou S. Application of change-point analysis to abnormal wind power data detection. Electric. Power Syst Res 2014;11:1807–11.
[63] Dasu T, Johnson T. Hunting of the shark: finding data glitches using data mining methods. Institute of Electrical Engineers; 1999.
[64] Yang Y, Li Y, He J. Detecting abnormal semantic web data using semantic dependency. Electric. Power Syst Res 2011;7:1546–55.
[65] Qiu C. The realization of small power plant data acquisition using the power data of four level network. Autom Electr Power Syst 2006;3:105–6.
[66] Shyh Jier H, Lin J. Enhancement of anomalous data mining in power system predicting-aided state estimation. IEEE Trans Power Syst 2004;1:610–9.

[67] Duc-Son P, Svetha V, Mihai L, Saha B. Anomaly detection in large-scale data stream networks. Data Min Knowl Discov 2014;9:281–5.

[68] Mao G, Duan L, Wang S. The principle and algorithm of data mining. Beijing: Tsinghua University Press; 2005.

[69] Sun G, Wei Z, Zhou F. Improved iterative self-organizing data analysis method of bad data identification. Chin J Electr Eng 2006;11:162–6.

[70] Mehmed K. Data mining: concepts, models, methods and algorithms. Beijing: Tsinghua University Press; 2003.

[71] Fasong W, Hongwei L, Rui L. Data mining with independent component analysis. IEEE Comput Appl Power 2004;3:6043–7.

[72] Zhou K, Yang S. Understanding household energy consumption behavior: the contribution of energy big data analytics. Renew Sustain Energy Rev 2016;56:810–9.

[73] Garcez R, Miranda V. Knowledge discovery in neural networks with application to transformer failure diagnosis. IEEE Trans Power Syst 2005;2:717–24.

[74] Rakhshani E, Sariri I, Rouzbehi K. Application of data mining on fault detection and prediction in boiler of power plant using artificial neural network. Int J Electr Power Energy Syst 2009:473–8.

[75] Teeuwsen SP. Neural network based multi-dimensional feature forecasting for bad data detection a vital and feature restoration in power system. IEEE Trans Power Appar Syst 2006;23:18–22.

[76] Souza JC, Leite , Silva AM, Alves , Silva AP. Information debugging in forecasting-aided state estimation using a pattern analysis approach. IEEE Trans Power Deliv 1996;2:123–32.

[77] Zhang G, Qiu J, Li J. Electric power data of the artificial neural network bad data identification and adjustment. Electr Eng 2001;8:104–7.

[78] Zhou K, Yang S. Exploring the uniform effect of FCM clustering: a data distribution perspective. Knowl-Based Syst 2016;96:76–83.

[79] Zhou KL, Fu C, Yang SL. Fuzziness parameter selection in fuzzy c-means: the perspective of cluster validation. Sci China Inf Sci 2014;57:1–8.

[80] Yue C. Decision theory and method. Beijing: China Science Press; 2003.

[81] Ozgur D, Murat T, Min A. An intelligent intrusion detection system (IDS) for anomaly and misuse detection in computer networks. Expert Syst Appl 2005;29:713–22.

[82] Richard F, Sue R, Lee B. Intrusion detection inter-component adaptive negotiation. Comput Netw 2001;34:605–21.

[83] Gao S, Shan Y. Advanced genetic algorithm approach to unit commitment with searching optimization. Proc CSEE 2001;3:45–8.

[84] Hu H. Application of parity mismatches on detection of bad data in power system state estimation. Procedia Eng 2011;15:198–207.

[85] Huang SJ, Lin J. Enhancement of power system data debugging using GSA based data-mining technique. IEEE Trans Power Syst 2002;11:1022–9.

[86] Feng L, Qiu J. Outlier data mining and its application in power load forecasting. Autom Electr Power Syst 2004;11:41–4.

[87] Wang C, Gao Y. Dynamic reconfiguration of distribution network based on optimal fuzzy C clustering and improved chemical reaction optimization. Proc CSEE 2014;10:1682–91.

[88] Tibshiriul R, Walther G, Hastie T. Estimating the number of clusters in a dataset via the gap Statistic. J R Stat Soc B 2000;63:411–23.

[89] Wu J, Yang W, Ge C. Elbow criterion for GSA power system based on bad data identification. In: Proceedings of the CSEE, 22; 2006. p. 23–8

[90] Yang W, Hu J, Wu J. The identification algorithm of bad data in power system based on GSA. Relay 2005;22:41–4.

[91] Zhang B. Application of GSA-based data mining for identifying bad data of power system. IEEE Trans Power Deliv 2003;18, [1007:115].

[92] Messina AR, Vittal V. A structural time series approach to modeling dynamic trends in power system data. Int J Electr Power Energy Syst 2012;9:1003–8.

[93] Guo Z, Li W, Adriel L. Detecting x-outliers in load curve data in power systems. IEEE Trans Power Syst 2012(2):875–84.

[94] Yan Y, Sheng G, Chen Y, Jiang X, Guo Z, Qin S. Big data cleaning method of time series analysis based on power transmission equipment. Autom Electr Power Syst 2015;7:138–44.

[95] Li Y, Li G, Tian ZH, Lu TB. A lightweight web server anomaly detection method based on transductive scheme and genetic algorithms. Comput Commun 2008:31.

[96] Su CT, Li GR. Reliability planning employing genetic algorithms for an electric power system. Appl Artif Intell 1999:13.

[97] Victor K, Denis S, Nikita T, Vadim S. Optimal training of artificial neural networks to forecast power system state variables. International. J Energy Optim Eng 2014:31.

[98] Xu NX, Seth D, Guikema Rachel A, Linda K, Çağnan Zehra, Kabeh V. Optimizing scheduling of post-earthquake electric power restoration tasks. Earthq Eng Struct Dyn 2006:362.

[99] Abdulwhab A, Billinton R, Eldamaty AA, Faried SO. Maintenance scheduling optimization using a genetic algorithm (GA) with a probabilistic fitness function. Electr Power Compon Syst 2004:3212.

[100] Dominik F, Thiemo G, Bernhard S. Swift rule: mining comprehensible classification rules for time series analysis, 5. . IEEE Trans on Knowledge and Data Engineering; 2011. p. 774–87.

[101] Zhan J, Liu K, Wang W, Liu Ying. Adaptive fuzzy clustering based anomaly data detection in energy system of steel industry. Inf Sci 2013;43:1908–11.

[102] Zhao J, Zhang G, Huang Y. With the present situation and prospect of new energy power system state estimation. Electr Power Autom Equip 2014;5:20–34.

[103] Huang S. Enhancement of anomalous data mining in power system predicting-aided state estimation. IEEE Trans Power Syst 2004(1):610–9.

[104] Zhang H, Li L. A hybrid approach for detection of bad data in power system state estimation. Power Syst Technol 2001;10:17–20.

[105] Huang G, Wu X, Yuan M. Research on metadata-driven data quality assessment architecture. Adv Neural Inf Process Syst 2013;67:4321–8.

[106] Guo C, Zhu C, Cao Y. Automation, research status and development trend of electric power system. Autom Electr Power Syst 2006;8:98–103.

[107] Xiao J, Zhang J, Zhu T. Study on electrical load correlation analysis based on city. Autom Electr Power Syst 2007;17:103–7.

[108] Cheng L, Li J, Liu J, Wen T. Correlation analysis of operation data and its application in operation optimization in power plant. Fuzzy Syst Knowl Discov 2008;18:581–5.

[109] Varun C, Arindam B, Vipin K. Anomaly detection. ACM Comput Surv (CSUR) 2009:3.

[110] Johanna H, Rocke David M. Outlier detection in the multiple cluster setting using the minimum covariance determinant estimator. Comput Stat Data Anal 2002:4.

[111] Thukaram D, Hadbanjan H, Phaniram Khincha M. Robust approach for identification of bad data in state estimation using SLP Technique. Int J Emerg Power Syst 2007:84.

[112] Murad A, Rassam Mohd, Aizaini Maarof, Anazida Zainal. Adaptive and online data anomaly detection for wireless sensor systems. Knowl Based Syst 2014;6:60–2.

[113] Zhang Y, Liu H, Zhou S. The innovation graph topology observability and bad data identification analysis. Autom Electr Power Syst 2008;6:55–9.

[114] Huang Q, Schulz N, Srivastava AK, Haupt T. Distributed state estimation with PMU using grid computing. Power Energy Soc Gen Meet 2009;45:1–7.

[115] Gao L, Wang H, Zhang W. Research on automatically clustering algorithm in web personalize service. Microelectron Comput 2007;12:1456–65.

[116] Jabr RA. Power system state estimation using an iteratively reweighted least squares method for sequential regression. Int J Electr Power Energy Syst 2006;2:86–92.