A Combined Algorithm for Data Cleaning of Wind Power Scatter Diagram Considering Actual Engineering Characteristics

JiaHui Wu¹, ZhenGuo Shao¹* and ShaoHua Yang¹
¹ College of Electrical Engineering and Automation, Fuzhou University, Fuzhou, Fujian, 350108, China
*Corresponding author’s e-mail: shao.zg@fzu.edu.cn

Abstract. There exist a lot of outliers in the wind power scatter diagram, which due to communication failure, equipment abnormality, artificial power limitation, and other reasons in the actual operation of wind turbines. These outliers have an adverse effect on performance and status evaluations for wind turbines seriously. Firstly, the reasonable power offset ranges are calculated using the affine algorithm, which considers actual engineering characteristics, e.g., forecast error and topographic difference. Besides, the quartering algorithm is applied for clearing the scattered outliers. Furthermore, the affine-quartile combined algorithm is presented to realize advantageous complementarities and attain accurate data of the wind power scatter diagram. Finally, Case studies show that the proposed affine-quartile algorithm, with consideration of actual engineering characteristics, can improve the coverage and accuracy of outlier detection.

1. Introduction

With the rapid development of the global economy, the demand for energy is increasing, the nonrenewable energy source is exhausted, as well as environmental pollution is becoming more serious year by year. Wind energy is a safe environment-friendly, renewable energy source. It is highly valued and becoming an important part of energy strategy in the world. In addition, wind power presents a good development prospect.

However, wind power generation is fluctuant and intermittent which brings great challenges to the stable operation of the power system. Data mining and power forecasting can improve the prediction of wind power output. However, there are a lot of incomplete and abnormal data in the original data, due to the reasons such as wind abandonment, maintenance, extreme weather conditions, external electromagnetic interference, and equipment failure during the operation of the wind farm. These situations will become more serious during the data transmission. Abnormal data affect the evaluation of the performance and operation of wind turbines seriously, which cannot provide data support for the economic and safe operation and optimization strategy of wind farms. Data cleaning can detect the outliers and missing value in the original data of wind turbines, then remove the outliers, reconstruct the missing value. This step can improve the data quality effectively, and is an important work in data preprocessing.

In recent years, many scholars at home and abroad are committed to the study of the cleaning method of wind power data, and they have made a lot of research results. At present, the most commonly used methods are mainly two kinds: one is to judge whether the data is abnormal according to the statistical
characteristics of the data itself, the other is to use the intelligent algorithm based on density or distance to detect outliers. In reference[1], the statistical analysis method is used to determine whether the wind power data is out of range, so the data which is less than or equal to zero will be eliminated. This method is simple, but it cannot deal with a large number of scattered outliers. Reference [2] assumed that the probability density function of wind power follows the normal distribution. Based on $3\sigma$ principles, the data that more than three times standard deviation is regarded as outliers. However, the actual wind power data is often not subject to the normal distribution strictly, resulting in the number of identified outliers is less than the actual outliers. In reference [3], the data is grouped by the quartile algorithm, and the range of outliers is obtained after calculating the internal limit, then the data beyond the internal limit will be eliminated. However, the internal limit can be affected by a large number of stacked outliers, resulting in poor recognition. In reference [4], support vector regression algorithm is used to fit the data, which can separate the outliers while smoothing the data, but this algorithm is slow in calculation speed and weak in robustness. In reference [5], k-nearest neighbor clustering algorithm is used to detect and separate outliers, but this algorithm needs a lot of data for training, otherwise, the training error will be large. Considering the variability of wind turbine operation state, it is difficult to obtain the normal data samples for training, so the practicability of this method is poor. In reference [6], the outlier detection algorithm is used, which uses a clustering method to identify outliers. But the ability of this method to identify stacked data with high density is poor. In reference[7], a combined algorithm of quartile-cluster algorithm is proposed to identify outliers. This combined algorithm can make up for the limitation of using a single algorithm. However, the influence of the cleaning sequence on the algorithm results is not considered.

In summary, the above two traditional data cleaning methods are used widely, but they ignore the actual engineering characteristics of wind turbines, and lack of pertinence to the wind power project. The results may lead to misidentification and a large deviation from the actual project data, which will greatly reduce the accuracy of cleaning. Given this, this paper proposes a new combined algorithm for wind power data cleaning. This new method combines the affine algorithm and the quartile algorithm. Firstly, the actual engineering characteristics of the wind turbines are considered in affine algorithm of data cleaning, which has a strong pertinence. After that, the affine-quartile algorithm is described. Besides, this paper verifies that the algorithms in different orders have different results. Further, this paper points out that compared with the other order, the affine-quartile algorithm can get a more accurate result of data cleaning. Finally, case studies show that the proposed affine-quartile algorithm is effective and reasonable.

2. Data Cleaning Method Based on affine Algorithm and Quartile Algorithm

2.1. Principle of affine algorithm

Data cleaning based on affine algorithm is divided into the following three steps:

1) According to the main influencing factors of wind speed, obtaining the range of its change by affine arithmetic:

The wind speed and the output power of the wind turbine will fluctuate in a certain range, which can be represented by the affine model. The uncertain sources of wind speed are meteorological factors and geographical factors, the most important of which are wind speed forecast errors and terrain differences. Therefore, wind speed prediction error and terrain difference are taken as noise elements when building the affine model of wind speed. In order to consider the practical engineering value and make the data cleaning more targeted, the noise element coefficient is obtained through the actual engineering data. Thus, the affine model of wind speed is as follows:

$$\hat{v}=v_0+x_1e_1+x_2e_2$$

where $v_0$ is the forecast wind speed, $e_1$, $e_2$ is the noise element introduced of the forecast wind speed and the terrain, and $x_1$, $x_2$ is the coefficient of the noise element. The standard deviation between the
calculated and measured wind speeds can be used as noise element coefficients $\varepsilon_1$, and the standard deviation between the average wind speed of each unit and the annual average wind speed of the whole wind farm can be used as noise element coefficients $\varepsilon_2$.

2) Fitting the center line curve of wind power scatter diagram:
The center line curve of wind power scatter diagram $f(v)$ is fitted by weighted least square method, which can reflect the center position of wind power data distribution without the influence of outliers. Then, the first derivative $f'(v)$ and second derivative $f''(v)$ of the $f(v)$ can be calculated for the subsequent Taylor expansion.

3) By considering the function mapping relationship of the physical characteristics of wind power generation, obtaining the range of power dispersion.

Taking the input wind speed range of wind farm as the independent variable, the affine function of wind speed power is established. Wind power can be expressed:

$$P(\hat{v}) = f(v_0 + x_1 \varepsilon_1 + x_2 \varepsilon_2)$$

(2)

$P(\hat{v})$ is the interval dispersion range of wind farm output power. By directly calculating affine $\hat{v}$ into affine $\hat{\hat{P}}(\hat{v})$, we can grasp the range of $\hat{\hat{P}}(\hat{v})$ change, and also retain the relationship between the uncertain influence factors and $\hat{\hat{P}}(\hat{v})$.

Then, the function is calculated based on the Taylor expansion and the uncertain fluctuation range of power is obtained, as follows:

$$\hat{\hat{P}}(\hat{v}) = f(v_0) + (x_1 \varepsilon_1 + x_2 \varepsilon_2) f'(v_0) + \frac{(x_1 \varepsilon_1 + x_2 \varepsilon_2)^2}{2} f''(v_0)$$

$$+ f(v_0) + \frac{|x_1| + |x_2|}{4} f'(v_0) + x_1 \varepsilon_1 f''(v_0) + x_2 \varepsilon_2 f''(v_0) + \frac{|x_1| + |x_2|}{4} f''(v_0) \varepsilon_3$$

(3)

$\varepsilon_3$ is the other factors that affect the wind speed error and the higher-order term of the original noise element in Taylor expansion. After finishing, the affine center value of power and the range of dispersion interval are obtained, and the edges of them are connected. The upper and lower envelope lines of wind speed power are obtained. The surrounding area is a strip area. The data points outside the area are considered as outliers that do not conform to the physical characteristics of wind power generation.

The first three steps of the affine algorithm get the interval dispersion range of wind power, forming a strip area. The outliers outside the envelope of the strip area can be eliminated.

Take 9.5m/s to 10m/s wind speed as an example, this data section contains communication error data, scattered data, and bottom stacked data. The distribution of outliers in this wind speed segment is shown in figure 1. The cleaning result by the affine algorithm is shown in figure 2.

Figure 1. The distribution of outliers
Figure 2. The Cleaning result by the affine algorithm

It can be seen from figure 2 that the affine algorithm eliminates the data that does not conform to the output law according to the physical characteristics of its output, including most of the around-curve outliers and bottom-curve stacked outliers. However, a small number of scattered data and top-curve outliers which close to the power curve have not been identified accurately. In theory, we can use the quartile algorithm to identify the remaining outliers.

2.2. The quartile algorithm
The quartile algorithm is also known as the box graph method. This algorithm groups the data through the median of statistics, and then achieve the data cleaning. The schematic diagram of the box type is shown in figure 3. In the quartile algorithm, the data is arranged from small to large and divided into four parts averagely, i.e. each series accounts for 25% of the whole series. It is worth mentioning that, $Q_1$ is called the lower quartile because that one-quarter of all data is smaller than the value of $Q_1$. Besides, $Q_2$ is the median and $Q_3$ is the upper quartile because that one-quarter of data is larger than the value of $Q_3$. The interquartile distance is the difference between the upper quartile and the lower quartile, which can be expressed as formula (4):

$$I_{QI} = Q_3 - Q_1$$

where $I_{QI}$ is interquartile distance.

According to the interquartile distance, the range of outliers can be obtained. Further, the data beyond the internal limit $[F_1, F_u]$ can be eliminated. The internal limit can be calculated by the formula (5):

$$[F_1, F_u] = [Q - 1.5I_{QI}, Q + 1.5I_{QI}]$$

Figure 3. Box diagram
The effect of using the quartile algorithm to clean the remaining outliers is shown in figure 4. It can be seen that the top-curve outliers and around-curve scattered outliers that are not recognized by affine arithmetic, but it can be effectively cleaned by the quartile algorithm. That is, the combination method of affine- quartile algorithm can identify four types of typical abnormal data effectively.

Figure 4. Effect picture of combined data cleaning method

2.3. The process by the combined algorithm

![Box diagram](image-url)
The affine algorithm can get rid of the outliers which in the middle of the curve, under the curve, and partial around the curve. But some around-curve scattered outliers cannot be identified when using the affine algorithm alone. The quartile algorithm can reject the remaining outliers. But this method can be affected by a large number of bottom-curve outliers. Therefore, the cleaning effect is poor when using the quartile algorithm only. The combination method does not need to train the data, and it makes up for the limitations of the single cleaning method. The physical characteristics of wind power generation are considered and the cleaning effect cannot be affected by bottom-curve outliers when using the combined method. This new method can eliminate a large number of outliers in a short time accurately, to improve the quality of wind power data and predictability of wind power generation.

3. Case study

3.1. Data sample
To verify the effectiveness of the new data cleaning method and its process, the data of wind turbines from a wind farm in Fujian province are taken as an example for the test. There exist 24 wind turbines on the farm with a rated power of 2MW. The cut-in wind speed, rated wind speed, and cut-out wind speed of the wind turbine are 3m/s, 15m/s, and 25m/s, respectively. The wind speed acquisition interval of the wind farm is 10min, and the acquisition time is from 00:00 on February 1, 2015 to 24:00 on July 31, 2015.

The distribution of data points is shown in figure 6. According to the distribution characteristics of data points, all four types of outliers exist, respectively are top-curve stacked outliers, mid-curve stacked outliers, around-curve scattered outliers, and bottom-curve stacked outliers. In the wind power scatter diagram, the bean method is used to draw the wind power curve.

Figure 6. The distribution of data points

3.2. Effect verification of the affine-quartile algorithm
The result of identifying outliers by the affine algorithm is shown in figure 7. And the result of identifying outliers by the affine-quartile algorithm is shown in figure 8.

Figure 7. The cleaning result by the affine algorithm
It can be seen from figure 8 that all four types of outliers in the wind turbine can be effectively identified. Therefore, the affine-quartile algorithm proposed in this paper, which considers actual engineering characteristics, is feasible to clean the outlier data of the wind turbine. Besides, advantages of the affine-quartile algorithm are summarized as follows.

1) The combination algorithm combines actual engineering characteristics of wind turbine generation, which can improve the pertinence of data cleaning to wind power engineering. The novel proposed affine-quartile algorithm has more engineering value and the phenomenon of wrong identification does not exist.

2) The low precision of the original affine algorithm can be compensated by the proposed affine-quartile algorithm. This proposed affine-quartile algorithm can detect and eliminate the scattered data close to the curve effectively.

3) The combination algorithm makes up for the defect that the quartile algorithm is easy to be affected by the bottom stacked outliers.

4) At the same time of improving cleaning precision, the new algorithm does not need a lot of data for training, which improves the speed of data cleaning greatly.

3.3. Comparative analysis of the different algorithms
In order to illustrate the rationality and effectiveness of the data cleaning process of the affine-quartile algorithm, this paper compares the new method with the quartile-affine algorithm and the most traditional quartile algorithm. The result of identifying outliers by the affine-quartile algorithm is shown in figure 9. And the result of identifying outliers by the quartile algorithm is shown in figure 10. The data deletion rate of three cleaning methods are shown in table 1.
As can be seen from Table 1, the data deletion rate by the affine-quartile algorithm, the quartile-affine algorithm, and the quartile algorithm are 10.2%, 7.99%, and 6.65%, respectively.

After adjusting the affine algorithm and the quartile algorithm in sequence, a large part of the around-curve scattered outliers has not been identified. It can be seen that when several cleaning methods are combined to process data, the advantages of the methods may be complementary or offset. The algorithms in different orders have different results. However, when the traditional quartile algorithm is used alone, due to the influence of a large number of bottom stacked data, the inner limit of abnormal value becomes larger, resulting in a large number of scattered outliers that cannot be recognized. Therefore, it can be proved that the new combined cleaning method is effective and reasonable.

In other words, under the condition of using an affine algorithm to eliminate a large number of bottom stacked data, the quartile algorithm can find out the accurate inner limit of outliers. Therefore, the combined algorithm cannot exchange the order, otherwise, the two algorithms cannot complement each other.

4. Conclusion
The wind's randomness and intermittence lead to the output of wind power not smoothly, so that there exist a lot of outliers in the original data.

At present, there are only mathematical methods and artificial intelligence methods for data cleaning of wind power, which do not take into account the engineering characteristics of wind turbines. This paper presents a new data cleaning method, a combined algorithm based on the affine-quartile algorithm.

The case studies show that the outliers in the wind speed power scatter diagram is accurately identified and eliminated by the proposed affine-quartile algorithm. Therefore, the novel combined cleaning method proposed in this paper has proved effective and reasonable. Besides, this algorithm has complementary advantages, i.e., it can solve the shortcomings of low accuracy of the affine algorithm and the weakness of easy to be influenced by bottom-curve stacked outliers.

Besides, the new algorithm improves the coverage and accuracy of data cleaning, as well as identified the four types outliers effectively. This novel algorithm takes the engineering characteristics of wind power generation into consideration so that the prediction and analysis results of wind turbine operation characteristics and operation status are more accurate and reliable.
Furthermore, the cleaning effect of outliers is not only related to the cleaning method, but also the cleaning sequence. Therefore, we further explore the combination of various cleaning methods and the impact of the use order on the data cleaning results.

In the future, the various combined cleaning methods and the influence of different order of cleaning methods are worthy of attention.

References

[1] Schlechtingen M, Santos I F. (2013) Achiche S. Using data-mining approaches for wind turbine power curve monitoring: a comparative study. IEEE Transactions on Sustainable Energy, 4(3): 671-679.

[2] Wang Y, Infield D G, Stephen B, et al. (2014) Copula-based model for wind turbine power curve outlier rejection. Wind Energy, 17(11): 1677-1688.

[3] Shen X, FU X, Zhou C. (2018) A combined algorithm for cleaning abnormal data of wind turbine power curve based on change point grouping algorithm and quartile algorithm. IEEE Transactions on Sustainable Energy, 10(1): 46-54.

[4] Cui H. (2018) Research on Eliminating Abnormal Big Data based on PSO-SVM. In: IEEE 3rd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), Chongqing. pp. 2460-2463.

[5] Kusiak A, Zheng H, Song Z. (2009) Models for monitoring wind farm power. Renewable Energy, 34(3): 583-590.

[6] Sun Z, Sun H. Stacked Denoising Autoencoder (2019) With Density-Grid Based Clustering Method for Detecting Outlier of Wind Turbine Components. IEEE Access, 7: 13078-13091.

[7] Dong A F, Zhang C H. (2017) Research on Abnormal Data Mining Algorithm. In: International Conference on Computer Systems, Electronics and Control (ICCSEC). Dalian. pp. 1477-1480.