Characteristic Curve Fitting Method of Wind Speed and Wind Turbine Output Based on Abnormal Data Cleaning

Bitong Han\textsuperscript{1}, Hongbin Xie\textsuperscript{1} and Yu Shan\textsuperscript{1} and Ruichen Liu\textsuperscript{*} and Shengxian Cao\textsuperscript{2}

\textsuperscript{1}State Grid New Energy Cloud Technology Co., Ltd., Beijing, 100000, China
\textsuperscript{2}School of Automation Engineering, Northeast Electric Power University, Jilin, 132012, China
*E-mail: liuruichen1212@outlook.com

Abstract. The characteristic curve of wind speed and power reflects the output state of wind turbine, and its characteristics are helpful to the accurate prediction of wind power. With the improvement of wind turbine power generation technology and other engineering applications, the data collected by SCADA system contains a large number of outliers, which makes it difficult to accurately fit the wind speed power curve. Firstly, this paper analyzes the types and causes of outliers in the actual data of wind turbines. Then, an abnormal data cleaning method based on Tukey’s method considering the operation parameters of wind turbines is proposed to clean the data of 12 wind turbines. Finally, the new data are clustered separately in the wind speed range by K-means, and the wind speed power characteristic curve is fitted. Compared with the uncleaned data, this method can significantly improve the fitting accuracy of wind speed power characteristic curve.

1. Introduction

With the requirements of countries for environmental quality, the proportion of new energy in energy structure is increasing [1]. At present, the main problem of wind power generation is how to stably connect to the grid and reduce the fault loss. To solve these problems, it is necessary to study effective methods. In the existing research and engineering examples, wind power prediction can provide effective reference significance for grid connection of wind turbines [2], and the identification of abnormal points in SCADA data can provide data support for early faults of wind turbines [3–4].

Wind power forecasting itself considers the relationship between wind speed and wind power, and the distribution of power points corresponding to different wind speed intervals can be observed intuitively by establishing wind speed power characteristic curves. Figure 1 shows the wind speed power curve drawn by the annual data of a wind farm in Iowa (93 26’ w, 40 52’ n) in 2012 (interval of 5 minutes). However, in the actual energy production, the data collected by SCADA will be abnormal due to the power limit caused by power grid dispatching, sensor collection errors and other reasons [5], which makes it difficult to fit the accurate wind speed power curve.

Therefore, this paper proposes a wind speed power curve fitting method combining Tukey’s method and K-means method. This method considers the parameters of the wind turbine itself, such as the cut-in wind speed and cut-out wind speed of the wind turbine (WT). Rated rotor speed, rated power, etc. By analyzing the overall distribution of the data, the Tukey’s method is used to clean the data. Finally, the cleaned data of each wind speed interval is clustered based on K-means method, and a more accurate wind speed power curve is fitted.
Figure 1. Characteristic curve of wind speed and power.

2. Analysis of Abnormal Value of Wind Speed-Power
A normal wind turbine should have a clean scatter diagram of wind speed and power, which means that the relationship between wind speed and power is relatively stable, and the abnormal data points will complicate the relationship between wind speed and power [6], which is unfavorable to the prediction of wind power and the research of wind speed and power curve. The data in figure 1 comes from a 1.5MW wind turbine in a wind farm in Northeast China. In this paper, the data of wind speed and power collected by SCADA system of wind turbine can be divided into four types.

Figure 2. Scatter chart of wind speed and power.

As shown in figure 2 and table 1, the Type I abnormal data points are mainly values with power at or near 0. It can be found that the data points in Type I are divided into left and right data accumulation zones by the WT cut-in speed of 2.5 m/s. The data accumulation band on the left side indicates the data point where the wind turbine is in pre-start or shutdown. The data accumulation
band on the right side indicates the point where the power of the wind turbine is still close to zero after
the wind speed reaches the cut-in wind speed, which indicates that the wind turbine may be in a state
of shutdown.

Table 1. Main types of abnormal points.

| Abnormal point type | Characteristic |
|---------------------|----------------|
| Type 1              | Exceeding the cut-in wind speed, the power value is less than 0 or close to 0. |
| Type 2              | When the cut-in wind speed is exceeded, the power value is low and the points are scattered. |
| Type 3              | In the lateral data point accumulation zone, when the wind speed is greater than the cut-in wind speed, the power value is basically constant. |
| Type 4              | The power value of this point is greater than the upper limit of normal power at this wind speed. |

Type II data points indicate that when the wind speed is high, the wind turbine does not produce
the normal power according to the wind speed. Most of these points are caused by the fact that the
wind turbine is in the final state of closing the propeller (close to shutdown), that is, the angle between
the blades of the wind turbine and the wind direction is too small, which makes the efficiency of
converting wind energy into electric energy of the wind turbine low.

Looking at the point of Type III, it can be clearly seen that this type of point can produce lateral
accumulation, which is about 15% lower than the case of full power generation. These points are due
to the forced air abandonment after the wind turbine is connected to the grid and receives the
dispatching instruction to meet the stable operation of the power grid, that is, the power limit point.

Finally, the power of data points of Type IV is 5% higher than that of corresponding wind speed,
which may be caused by the failure of wind speed sensor.

3. Data Cleaning
Section 2 describes four types of outliers in SCADA data. For these four types of outliers, we first use
the threshold of WT parameters to clean them in the first step, and then use Tukey’s method to process
the outliers of the whole data.

3.1. Conventional Method of Considering WT Parameters
The data is selected from the open data set of China’s 2020 Green Future-Wind Power Abnormal Data
Identification and Cleaning Competition. The data has the data of 12 wind turbines for one year, and
the parameters include wind speed \( v_{\text{wind}} \), wind power \( P_{\text{wind}} \), and rotor speed \( v_{\text{rotor}} \). The model number
of each typhoon is different, and the racing competition party gives the parameters of each WT,
including the diameter of the wind wheel, rated power \( P_{\text{rated}} \), cut-in wind speed \( v_{\text{cut-in}} \), cut-out wind
speed \( v_{\text{cut-out}} \), and the range of wind wheel speed \( R_{\text{rotor speed}} \).

Algorithm 1: Threshold limit method

\[
\begin{align*}
\text{Data: } & v_{\text{wind}}, P_{\text{wind}}, v_{\text{rotor}} \text{ of 12 wind turbines} \\
\text{Conditions: } & v_{\text{cut-in}}, v_{\text{cut-out}}, P_{\text{rated}}, R_{\text{rotor speed}} \text{ of 12 wind turbines} \\
\text{Result: } & v_{\text{wind}}, P_{\text{wind}}, v_{\text{rotor}} \text{ falling within a given range} \\
1 & \text{Select the data of a wind turbine and check whether the parameters meet the conditions.} \\
2 & \text{while } v_{\text{cut-in}} < v_{\text{wind}} < v_{\text{cut-out}} \text{ and } 0 < P_{\text{wind}} < P_{\text{rated}} \text{ and } v_{\text{rotor}} \in R_{\text{rotor speed}} \text{ do} \\
3 & \text{Use ‘0’ to mark this data as normal.}
\end{align*}
\]

As shown in Algorithm 1, firstly, according to the given WT parameters, this paper uses threshold
limit method to filter the data of each WT. After processing, the normal parts of the data are marked as
‘0’.
3.2. Outliers Processing based on Tukey’s Method

Tukey’s method is a classic outlier processing method, which calculates the inner distance IQR (inter-quartile range) of the sample to get the inner limit of the whole sample. The values falling within the inner limit are normal values and those falling within the outer limit are abnormal values [7]. This method is stricter than Pauta Criterion in detecting outliers, and it doesn’t require the data to conform to normal distribution like the latter criterion.

![Figure 3. Schematic diagram of Tukey’s method.](image)

As shown in figure 3, when the values of wind speed, wind power and rotor speed of a sample can all fall within their respective inner limits, the sample is normal, and the sample is marked with ‘0’, otherwise, it is abnormal and marked with ‘1’.

Tukey’s method can handle every parameter with exception, which effectively reduces the probability of missing abnormal samples. It is worth mentioning that the data used in this paper only contains three types of data, and when the data set with more parameters is encountered, the abnormal samples found are more accurate.

4. Fitting of Characteristic Curve

The reliable fitting method is beneficial to improve the accuracy of the characteristic curve. In this paper, a K-means method for clustering power points under different wind speeds is used to fit the characteristic curve of wind speed-power. K-means method is an unsupervised learning clustering algorithm with no sample output. It can find the best centroid of k categories through training, and then determine the cluster category of the sample [8]. For the characteristic curve fitting problem, we divide the wind speed into several intervals at intervals of 1m/s, and each interval needs to find a point as one of the components of the characteristic curve.

![Figure 4. Schematic diagram of K-means.](image)

As shown in figure 4, use K-means to cluster points in each wind speed interval, where K is 1, and find a cluster center. Finally, find out the points of N wind speed intervals, and finally get the required wind speed-power characteristic curve by connecting the points.
5. Experiment

5.1. Introduction of Data
The given conditions of the data are shown in table 2. There are 497,838 pieces of data for 12 wind turbines, with an average of 41,486 pieces for each wind turbine, and each sample is sampled at 10-minute intervals.

| IDWT | $P_{rot}$ (kW) | $v_{cut,in}$ (m/s) | $v_{cut,out}$ (m/s) | $R_{rot}$ speed (r/min) |
|------|----------------|-------------------|---------------------|-------------------------|
| 1    | 2000           | 3                 | 25                  | 8.33-16.8               |
| 2    | 2000           | 3                 | 25                  | 8.33-16.8               |
| 3    | 2000           | 3                 | 25                  | 8.33-16.8               |
| 4    | 2000           | 3                 | 25                  | 8.33-16.8               |
| 5    | 2000           | 3                 | 22                  | 5.5-19                  |
| 6    | 2000           | 3                 | 25                  | 8.33-16.8               |
| 7    | 2000           | 3                 | 25                  | 8.33-16.8               |
| 8    | 2000           | 3                 | 25                  | 8.33-16.8               |
| 9    | 2000           | 3                 | 25                  | 8.33-16.8               |
| 10   | 2000           | 3                 | 25                  | 8.33-16.8               |
| 11   | 2000           | 2.5               | 19                  | 5-14                    |
| 12   | 2000           | 3                 | 22                  | 5.5-17                  |

5.2. Experiment 1: Data Cleaning
As shown in figure 5, we selected the data of No.11 WT as the main research object. From its original data, the location of its outliers is similar to that contained in figure 2 in the Section 2. Through threshold limit and Tukey method processing, it can be clearly seen that the data points of Type I, II, III and IV in the ‘Health data’ part of the figure can basically be completely eliminated, which provides an advantage for fitting accurate wind speed-power characteristic curve in the next step.

Figure 5. Data cleaning result of No.11 WT.
5.3. Experiment 2: Characteristic Curve Fitting
K-means method is used to cluster the wind speed-power points before and after data cleaning with \( k = 1 \). Thus, two sets of cluster center points with different lengths are obtained. Using the wind speed and each cluster center point, we can get the comparison chart of wind speed-power characteristic curve as shown in figure 6.

![Figure 6. Comparison of fitting results of characteristic curves.](image)

It can be seen from figure 6 that the two wind speed power curves can be well fitted at low to medium wind speeds. However, we can’t see the obvious difference between the two curves at the middle and low wind speeds well only by visual observation. Using the error between the two curves to describe the difference.

| Evaluating indicator | MAE | RMSE | MAPE | \( R^2 \) |
|----------------------|-----|------|------|---------|
| Value                | 5.44| 7.19 | 0.27%| 0.99    |

As shown in table 3, take the first 14 points of two curves and calculate the MAE (Mean Absolute Error), RMSE (Root Mean Square Error), MAPE (Mean Absolute Percentage Error) and \( R^2 \) (R Square) between the two curves. In addition, at the wind speed of 12.5m/s, the characteristic curve that has not been cleaned by the data is affected by the abnormal values in the data, resulting in the power point not conforming to the actual output characteristics.

\[
MAE = \frac{1}{n} \sum_{i=1}^{n} |\hat{Y}_i - Y_i| \tag{1}
\]

\[
MAPE = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{\hat{Y}_i - Y_i}{Y_i} \right| \times 100\% \tag{2}
\]

\[
RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (\hat{Y}_i - Y_i)^2} \tag{3}
\]

\[
R^2 = \frac{\sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2} \tag{4}
\]
where, $\hat{Y}_i$ represents the value after data cleaning, $Y_i$ represents the value of no data cleaning, $\bar{Y}_i$ represents the mean value of $Y_i$.

### 6. Conclusion
In this paper, a SCADA data cleaning method combining threshold limit and Tukey method is proposed. Based on the data processed by this method and K-means clustering, a more accurate wind speed-power characteristic curve can be obtained. Firstly, the threshold limit method is used to filter the outliers that do not meet the characteristics of wind power production. Then, the abnormal values of each parameter sequence in the data are searched by Tukey method. It is found that there is a strong relationship between wind speed, power and rotor speed, and abnormal points can usually be found in pairwise comparison of parameters. This shows that the more parameters in the data, the smaller the probability of missing abnormal data. Finally, K-means is used to find the cluster center points in each wind speed partition, and the wind speed power curve is obtained. The experimental results show that K-means can better find the cluster center that meets the characteristics in the data cleaned, and further improve the fitting accuracy of wind speed power characteristic curve.

### Acknowledgements
This Paper was supported by Research on Blood Relationship Tracking and Automatic Error Correction Technology of State Grid New Energy Cloud Data Based on Probe Technology (SGDS0000WLJS2100131).

### References
[1] Chen L, Li K, Chen S, et al. 2021 Industrial Activity, Energy Structure, and Environmental Pollution in China Energy Economics 104 105633
[2] Li J, Liu C, Zhang P, et al. 2020 Difference between grid connections of large-scale wind power and conventional synchronous generation Global Energy Interconnection 3 486-493
[3] Zhang S and Lang Z 2020 Scada-data-based wind turbine fault detection: a dynamic model sensor method Control Engineering Practice 102 104546
[4] Xiang L, Wang P, Yang X, et al. 2021 Fault detection of wind turbine based on Scada data analysis using CNN and LSTM with attention mechanism Measurement 175 109094
[5] Dong F and Shi L 2019 Regional differences study of renewable energy performance: a case of wind power in China Journal of Cleaner Production 233 490-500
[6] Chen H, Liu H, Chu X, et al. 2021 Anomaly detection and critical scada parameters identification for wind turbines based on LSTM-ae neural network Renewable Energy 172 829-840
[7] Smiti A 2020 A critical overview of outlier detection methods Computer Science Review 38 100306
[8] Fahim A 2021 K and starting means for K-means algorithm Journal of Computational Science 55 101445