Data-Driven Correction Approach to Refine Power Curve of Wind Farm Under Wind Curtailment

Article *in* IEEE Transactions on Sustainable Energy · June 2017

DOI: 10.1109/TSTE.2017.2717021

6 authors, including:

Yongning Zhao
Cardiff University
23 PUBLICATIONS **285** CITATIONS
SEE PROFILE

Yuntao Ju
Tsinghua University
31 PUBLICATIONS **267** CITATIONS
SEE PROFILE

Data-Driven Correction Approach to Refine Power Curve of Wind Farm Under Wind Curtailment

Yongning Zhao, *Student Member, IEEE*, Lin Ye, *Senior Member, IEEE*, Weisheng Wang, *Senior Member, IEEE*, Huadong Sun, *Senior Member, IEEE*, Yuntao Ju, *Member, IEEE*, and Yong Tang, *Senior Member, IEEE*

***Abstract*—Regardless of the rapid development of wind power capacity installation around the world, wind curtailment is a severe problem to be solved. Wind curtailment can cause abundant outliers and change the original characteristics of operation data in wind farms. Power curve cannot be accurately modeled with these outliers and consequently wind power forecasting as well as other applications in power system will be negatively affected. In this paper, the characteristics of the outliers caused by wind curtailment are analyzed. Then, a data-driven outlier elimination approach combining quartile method and density-based clustering method is proposed. First, the quartile method is used twice for eliminating sparse outliers. Then density-based spatial clustering of applications with noise method is applied to eliminate stacked outliers. A case study is carried out by modeling the power curves of a wind farm and 20 wind turbines in this wind farm. The accuracy of power curve modeling is significantly improved and the elimination procedure can be completed in a very short time, indicating that the proposed methods are effective and efficient for eliminating outliers. The performance of the methods is insensitive to their parameters and can be directly used in different cases without tuning parameters, both for wind turbines and wind farms.**

***Index Terms*—Outlier elimination, power curve, wind curtailment, wind farm modeling, wind power.**

I. Introduction

WIND power has seen its increasingly important role in mitigating environmental problems and energy crisis, which makes it the most rapidly growing renewable energy over last decades [1], [2]. As wind penetration levels increase, wind farm operation data especially wind speed and wind power data are crucial and have been widely used in wind integration studies [3], [4] to help develop more advanced methods for accommodating uncertainties. Notably, the power curves [5]–[7] derived from synchronous wind speed and wind power data are indispensable in numerous applications. For example, they are frequently employed for monitoring the performance and operation conditions of wind turbines [8], [9] and wind farms [10], [11]. Trained with historical data, the power curve model is then applied using the wind speed as input to obtain an expectation of the power output. The conditions of wind turbine or wind farm can be detected by comparing the real operating conditions with the expectation. Other significant applications of power curves in power systems are wind power forecasting [12], [13], wind resource assessment [14], [15], wind turbine site matching [16] and power system reliability assessment [17], [18], etc.

Modeling accurate power curves requires uncontaminated data collected from practical operation of wind turbines, which is essential for characterizing the true statistical relationship between wind speed and wind power output [19]. However, the fact is that wind turbines generally operate in the environments with large uncertainties so that many outliers could be interfused into the operation datasets. The outliers can be caused by storage and communication errors, wind turbine malfunctions, icing on blades and wind curtailment, which may not be well documented [20], [21]. The presence of outliers will damage the data quality and lead to inaccurate wind power curve modeling that may subsequently affect power system operation. In particular, severe wind curtailment brings abundant outliers into the data recordings. The outliers caused by wind curtailment present particular shapes in the wind speed-power (v-p) scatter plot. They are distributed horizontally along the horizontal axis (i.e., the wind speed axis) and form data clusters or data bars, which will badly distort the original statistical relationship between wind speed and wind power. This kind of outliers is called stacked outliers in this paper, while the outliers caused by random noise are sparse outliers.

Recently, the outliers have attracted much attention in the field of power curve modeling. Kusiak *et al.* [10] filtered out outliers by using residual approach and control charts. A k-NN method was used in advance to extract normal power curve from original data with outliers. But obviously they didn't consider the case with more outliers, for which the k-NN would not be applicable. Wan *et al.* [22] presented a kind of vertical stacked outliers, which implies that the wind farm produces power higher than theoretical value at a given low wind speed. This kind of outliers can be removed by simple statistical

Manuscript received October 30, 2016; revised March 13, 2017; accepted June 7, 2017. Date of publication June 19, 2017; date of current version December 14, 2017. This work was supported in part by the National Natural Science Foundation of China under Contracts 51477174, 51677188, and 51711530227, in part by the National Key Research and Development Program of China under Grant 2017YFB0902200, in part by the Project of State Grids Corporation of China under Contract 5201011600TS, and in part by the Open Fund of State Key Laboratory of Operation and Control of Renewable Energy and Storage Systems. Paper no. TSTE-00833-2016. *(Corresponding author: Lin Ye.)*

Y. Zhao, L. Ye, and Y. Ju are with the College of Information and Electrical Engineering, China Agricultural University, Beijing 100083, China (e-mail: zyn@cau.edu.cn; yelin@cau.edu.cn; juyuntao@cau.edu.cn).

W. Wang, H. Sun, and Y. Tang are with the China Electric Power Research Institute, Beijing 100192, China (e-mail: wangws@epri.sgcc.com.cn; sunhd@epri.sgcc.com.cn; tangyong@epri.sgcc.com.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSTE.2017.2717021

methods. In [23] a copula-based joint probability model was proposed for wind turbine power curve outlier rejection. However, their model only worked for sparse outliers while the stacked outliers were still remained in the data. Mangalova *et al.* [24] detected and eliminated outliers by using a parameterized formula to improve the data quality and wind power forecasting accuracy. The method is simple to implement, but the parameters have to be carefully tuned to adapt to different cases. Ye *et al.* [21] described an outlier identification method based on probabilistic wind farm power curve and typical outlier characteristics. But this method requires the outliers to be temporally consecutive. Villanueva and Feijóo [25] proposed a model for true power curve by fitting the wind power in each range of wind speed into a normal probability distribution. The data beyond the range of three times standard deviations are identified as outliers. However it is difficult to satisfy the assumption of normal distribution in some cases, such as the topic of abundant outliers discussed in the following section. Park *et al.* [20] developed an automatic power curve limit calculation algorithm to exclude outliers by iteratively calculating average and standard deviation of power values per each speed interval. But the average and standard deviation are sensitive to abundant outliers which can make this method fail to work.

For most of the outlier identification methods in existing literature, the following condition needs to be satisfied: The outliers show a small effect on the power curve modeling. These methods explicitly or implicitly assume that the outliers will not significantly change statistical characteristics of original data, such as average, standard deviation and probability distribution. However, this condition will not hold for the abundant stacked outliers caused by wind curtailment due to their special characteristics, which will be explained in more details in the rest of the paper.

In this paper, based on the characteristics analysis of the *v-p* scatters, a combined approach of the quartile method and density-based cluster method is proposed to identify and eliminate the abundant outliers caused by wind curtailment. Neither any specification from wind turbine manufacture nor statistical assumption on the data is required to implement the proposed approach. It is a totally data-driven and unsupervised learning approach. The performance of proposed methods is insensitive to their parameters and can be directly used in different cases, for both wind turbines and wind farms without carefully tuning the parameters. The outlier elimination methods are tested on real operational data and proved to be effective and efficient.

The characteristics of outliers caused by wind curtailment and their impact on the power curve modeling are outlined in Section II. The proposed outlier elimination methods are introduced in Section III. A case study is carried out in Section IV to present the detailed implementation of proposed methods. Conclusions are given in Section V.

II. The Characteristics of Outliers Caused by Wind Curtailment

The relationship between wind speed v at hub height and wind power output P of a wind turbine can be characterized as a power curve model:

$$P = \begin{cases} 0 & v < v_{\mathrm{i}}, v > v_{\mathrm{o}} \\ P(v) & v_{\mathrm{i}} \le v \le v_{\mathrm{r}} \\ P_{\mathrm{r}} & v_{\mathrm{r}} \le v \le v_{\mathrm{o}} \end{cases} \tag{1}$$

where v_{i} is the cut-in speed at which the wind turbine starts to generate power; v_{o} is the cut-out speed at which the wind turbine stops generating power for mechanical protection; v_{r} is the rated speed at which wind turbine generates its rated power of P_{r}.

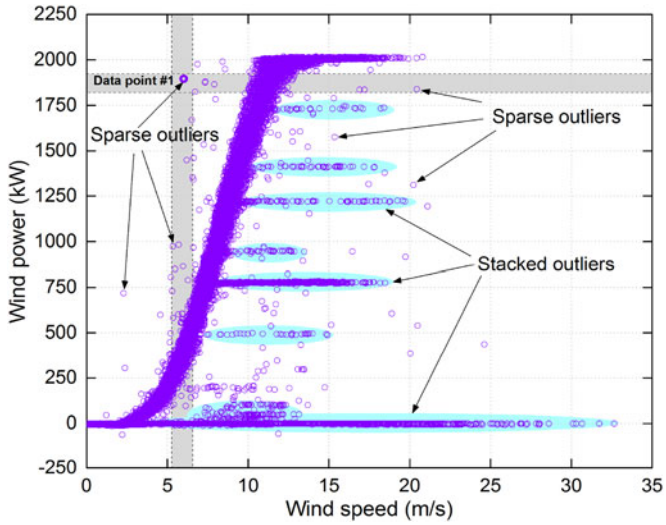


Fig. 1. The characteristics of outliers in the *v-p* scatters.

The power curve model given by the manufacture of a wind turbine is measured under predefined standard conditions [26]. In practical operation, the conditions will differ from standard ones. Therefore, it is necessary to extract practical power curve model from operation data. One of the most difficulties is that the operation data always contains outliers, which have to be removed before power curve modeling. In this paper, the focus is on analyzing two kinds of outliers, i.e., sparse outliers and stacked outliers, which are depicted in Fig. 1.

The number of sparse outliers is small. They are decentralized and sparsely distributed in the whole region of the scatter plot. For the convenience of the description of proposed methods in Section III, the sparse outliers are further defined in two situations: horizontal sparse outliers and vertical sparse outliers. Taking the data point #1 in Fig. 1 as example, it is called horizontal sparse outlier when investigated in the background of a wind power interval, i.e., the horizontal grey rectangle, whereas it is called vertical sparse outlier when investigated in a wind speed interval, i.e., the vertical grey rectangle.

Compared with sparse outliers, the stacked outliers caused by wind curtailment are horizontally and densely distributed in the scatter plot. Wind curtailment is defined as a practical operation by manual intervention to suppress and reduce the wind power outputs of wind generation units, which is intended to keep the security and stability of power systems operation. Wind curtailment can be triggered by many factors, such as transmission congestion, wind power ramping, underestimated

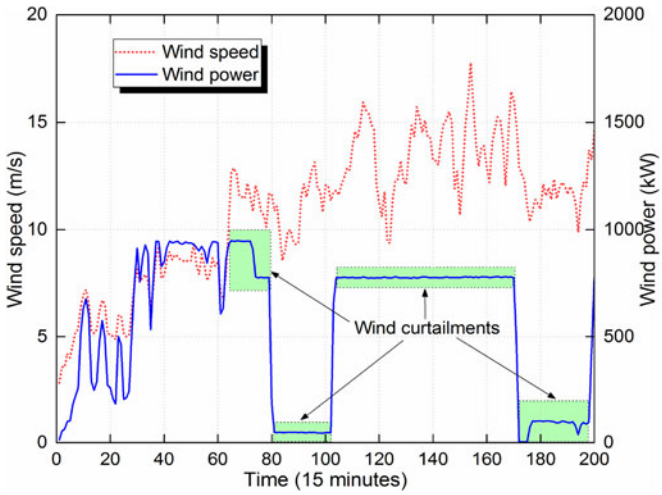


Fig. 2. The stacked outliers presented in the form of time series.

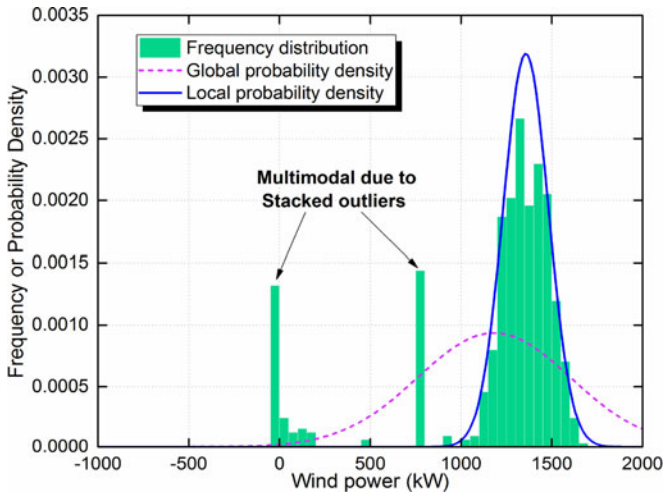


Fig. 3. The wind power distributions for a selected bin of a wind turbine.

wind power generation and excess generation relative to load levels [27], [28]. When a curtailment occurs, the wind turbine will produce lower power than the theoretical value defined by (1).

The stacked outliers in Fig. 1 can also be illustrated in the time series plot. Fig. 2 shows time series of wind speed input and wind power output of a wind turbine, of which v_{i} is 3.5 m/s, v_{r} is 12 m/s, v_{o} is 25 m/s and P_{r} is 2 MW. In Fig. 2, when 60< t <200, the wind speeds are around or higher than the rated wind speed. The outputs of the wind turbine for these wind speed values are supposed to reach the rated power of 2000 kW. However, the wind power output is curtailed to different lower levels than the theoretical level. When 100< t <170, during which most of the wind speed values are higher than 12 m/s, the output power is reduced to 780 kW. For 80< t <100 and 170< t <200, the power almost drops to zero.

To give a deep insight into the characteristics of the stacked outliers, the scatter plot is mapped to probability distributions. According to the standard of IEC 61400-12-1 [26], the wind speed is divided into a number of bins or intervals with the same width. The width of the bins is dependent on specific cases. In this paper, the width is set as 0.5 m/s. A typical interval of (9.0, 9.5] m/s is chosen for analysis. The frequency distribution histogram of wind power in this interval is shown in Fig. 3. It can be seen that the wind power present significant multimodal distribution with an outlined quasi-Gaussian distribution on the right, while two distinct isolated peaks are identified on the left. The peaks are in accordance with the horizontally distributed stacked outliers in Fig. 1, revealing that the multimodal feature is caused by stacked outliers. A global probability density function of normal distribution is fitted to the wind power in the bin, whereas a local probability density function is fitted to the right part of data in Fig. 3 by excluding the left part of data. The global probability density function is flat while the local one is steep. The stacked outliers have a remarkable effect on characterizing the probability distribution of wind power. In other words, the power curve model cannot be accurately modeled from scatter plot with these outliers.

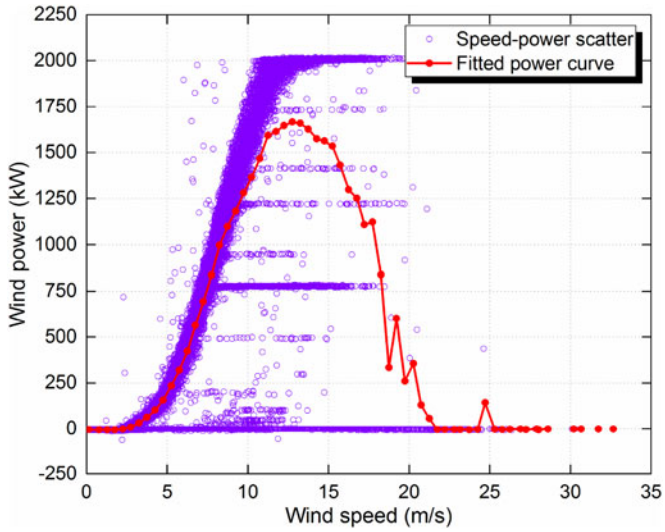


Fig. 4. The power curve modeled by 'bin' method using data with stacked outliers.

The power curve model is obtained using 'bins' method and given in Fig. 4. The fitted power curve significantly deviates from the central part of the scatters, especially when the wind speed is higher than 9 m/s. That indicates the power curve modeled from data with stacked outliers is inaccurate. The outliers have to be eliminated before the data are put into use, e.g., power curve modeling, wind turbine monitoring, etc.

III. The Data-Driven Approach for Outlier Elimination

A. Quartile Method for Eliminating Sparse Outliers

The quartiles are three values that divide an ordered dataset into four equal parts [29]. Each part contains 25% of the total observations in the dataset.

In this paper, the dataset to be investigated is either the wind power observations in a wind speed interval or the wind speed observations in a wind power interval. Both of the two cases share the same principal of quartile method. Taking the former as example, for the dataset in a wind speed interval

comprising n wind power observations in ascending order, $\boldsymbol{P}_v = \{p_1, p_2, \ldots, P_n\}$, the following method [30] is applied to compute the quartiles.

1) Calculate the second quartile P_2, i.e., the median.

$$P_2 = \begin{cases} p_{\frac{n+1}{2}} & n = 2k+1; k = 0, 1, 2, \ldots \\ 0.5p_{\frac{n}{2}} + 0.5p_{\frac{n+2}{2}} & n = 2\,k; k = 1, 2, \ldots \end{cases} \tag{2}$$

2) Calculate the first quartile P_1 and third quartile P_3.

If $n = 2\,k\,(k = 1, 2, 3, \ldots)$, split $\boldsymbol{P}_v$ into two parts by P_2. P_2 is not included in the two parts. Calculate the second quartiles of the two parts respectively and denote them as P_2' and P_2'' ($P_2' < P_2''$). Then $P_1 = P_2'$, $P_3 = P_2''$.

If $n = 4\,k + 3\,(k = 0, 1, 2, \ldots)$,

$$\begin{cases} P_1 = 0.75p_{k+1} + 0.25p_{k+2} \\ P_3 = 0.25p_{3k+2} + 0.75p_{3k+3} \end{cases}. \tag{3}$$

If $n = 4\,k + 1\,(k = 0, 1, 2, \ldots)$,

$$\begin{cases} P_1 = 0.25p_k + 0.75p_{k+1} \\ P_3 = 0.75p_{3k+1} + 0.25p_{3k+2} \end{cases}. \tag{4}$$

After the calculation of P_1 and P_3, the Interquartile Range (*IQR*) can be obtained by

$$IQR = P_3 - P_1. \tag{5}$$

The normal data in $\boldsymbol{P}_v$ are defined by two limits

$$[F_l, F_u] = [P_1 - 1.5IQR, P_3 + 1.5IQR] \tag{6}$$

where the F_l is lower limit and F_u is upper limit. The wind power observations that fall in $[F_l, F_u]$ are recognized as normal ones, whereas outliers are observations that fall below F_l or above F_u.

The *IQR*, F_l and F_u are robust measures of the distribution of a dataset, because they are almost unaffected by some individual outliers. The quartile method can be easily implemented to detect and eliminate outliers. But it only works well for the dataset with a small number of outliers. In this paper, the number of stacked outliers is comparable with or even larger than that of normal data. In this case, the quartile method is ineffective to detect the stacked outliers. Therefore, the quartile method is only used to eliminate the sparse outliers. The stacked outliers are left to be dealt with using the density-based clustering method. As shown in Fig. 1, sparse outliers distribute among the stacked outliers. Elimination of these sparse outliers will make the boundaries of the stacked outliers clearer. This is helpful to the clustering of stacked outliers.

B. Density-Based Clustering Method for Eliminating Stacked Outliers

Clustering is one of the most important unsupervised learning methods. It is the classification of similar objects into different groups (clusters), so that the data in each cluster show similar features while data in different clusters present great difference. Many algorithms have been developed for cluster analysis, among which the most popular ones are *k*-means and hierarchical clustering methods [31]. As has been shown in Fig. 1, the number of clusters formed by stacked outliers is different for various wind speed intervals and wind turbines. This number is unknown and difficult to determine in advance. Therefore, clustering methods like *k*-means are not suitable here.

According to the characteristics of the stacked outliers, the DBSCAN (density-based spatial clustering of applications with noise) method [32] is introduced in this paper. The key idea is that for each point of a cluster the neighborhood of a given radius has to contain at least a minimum number (*MinPts*) of points, i.e., the density in the neighborhood has to exceed a certain threshold. For the dataset in a wind speed interval comprising n wind power observations, $\boldsymbol{P}_v = \{p_1, p_2, \ldots, p_n\}$, the following rules are involved in DBSCAN method.

1) The neighborhood of a point p_i, denoted by $\boldsymbol{N}_\varepsilon(p_i)$ is defined by $\boldsymbol{N}_\varepsilon(p_i) = \{p_j \in \boldsymbol{P}_v | \mathrm{dist}(p_i, p_j) \le \varepsilon\}$, where the *dist*($p_i$, p_j) is any kind of distance function for two points.
2) A point p_j is directly density-reachable from a point p_i with regard to ε and *MinPts* if $p_j \in \boldsymbol{N}_\varepsilon(p_i)$ and the points in $\boldsymbol{N}_\varepsilon(p_i)$ are more than *MinPts*. p_i is called core point.
3) A point p_j is density-reachable from a point p_i with regard to ε and *MinPts* if there is a chain of points $p_i, p_{i+1}, \ldots, p_j$ such that p_{i+1} is directly density-reachable from p_i.
4) A point p_j is density-connected to a point p_i if there is a point p_k such that both, p_i and p_j are density-reachable from p_k.
5) A cluster $\boldsymbol{C}$ with regard to ε and *MinPts* is a non-empty subset of $\boldsymbol{P}_v$ satisfying the following conditions:
   a) $\forall\ p_i, p_j$: $p_i \in \boldsymbol{C}$ and p_j is density-reachable from p_i then $p_j \in \boldsymbol{C}$.
   b) $\forall\ p_i, p_j \in \boldsymbol{C}$: p_j is density-connected to p_i.
6) The noise is defined as the set of points in $\boldsymbol{P}_v$ not belonging to any cluster.

Notice that though DBSCAN has the ability to identify the noise (i.e., the sparse outliers), it is not reliable when applied for sparse outlier elimination in this paper. Instead, the quartile method is used, for which the reason is further explained in Section IV.

To find a cluster, DBSCAN starts with an arbitrary point p_i and retrieves its ε-neighborhood. If p_i is a core point, then a cluster $\boldsymbol{C}$ including p_i is generated. All density-connected points of p_i are added to *C*. Then a new unvisited point will be retrieved to find a new cluster by repeating the above process. The procedure of DBSCAN method is summarized as follows.

1) Set the values of ε and *MinPts*.
2) Arbitrarily select an unvisited point p_i and mark it as visited. If p_i is a core point, find its ε-neighborhood $\boldsymbol{N}$ and generate a new cluster $\boldsymbol{C}$ including p_i. Otherwise, mark it as noise.
3) Retrieve each point p_j in $\boldsymbol{N}$; if p_j is unvisited, mark it as visited; if p_j doesn't belong to any cluster, add it to $\boldsymbol{C}$; if p_j is a core point, merge the ε-neighborhood of p_j into $\boldsymbol{N}$.
4) Repeat Step 3) until no more core points are found. Then return to Step 2).

DBSCAN method is efficient and does not require one to specify the number of clusters in the dataset a priori, as opposed to *k*-means. It only requires two parameters: ε and the minimum

number of points required to form a dense region (*MinPts*). For the case in this paper, the outlier elimination performance is not sensitive to the two parameters. Thus DBSCAN is very suitable for elimination of stacked outliers.

C. The Procedure of Outliers' Elimination Using Quartile Method and DBSCAN

1) Preliminary Elimination: In normal operation condition, it is unreasonable for a generator to yield negative wind power, which can be seen as outliers. So the wind power that smaller than zero and their corresponding wind speed are eliminated from the scatter plot.

2) The Elimination of Horizontal Sparse Outliers Using Quartile Method: Firstly, the *v-p* data pairs are sorted by wind power in ascending order. Then wind power values are divided into some equal intervals. The quartile method is applied to the wind speed dataset in each power interval. The wind speed data beyond $[F_l, F_u]$ are eliminated from the dataset.

3) The Elimination of Vertical Sparse Outliers Using Quartile Method: The *v-p* data pairs are sorted by wind speed in ascending order. Then the wind speed values are divided into a number of equal intervals. The quartile method is applied to the wind power dataset in each wind speed interval.

However, the elimination of vertical sparse outliers is different from that of horizontal ones. Only the wind power data above F_u are eliminated from the dataset while the data below F_l are not considered. On one hand, as has been mentioned, the quartile method only works for the dataset with a small number of outliers. However for this case, in each wind speed interval, the outliers in the lower part may be more than normal data in the upper part. This leads to very few or even no data below lower limit F_l. It is useless and unnecessary to consider F_l. On the other hand, the elimination of vertical sparse outliers in each wind speed interval is intended for the topmost outliers that the horizontal elimination cannot deal with. This is very important because the DBSCAN method works by eliminating all the clusters except for the topmost one. Thus it is priority to ensure the topmost scatters are not outliers.

4) The Elimination of Stacked Outliers Using DBSCAN: The *v-p* data pairs are sorted by wind speed in ascending order. Then the wind speed values are divided into a number of equal intervals. The DBSCAN clustering method is applied to the wind power dataset in each wind speed interval.

It can be inferred from Fig. 1 that the stacked outliers are supposed to be in the lower part of each wind speed interval. Therefore, among the partitioned clusters in a wind speed interval, the topmost cluster with largest average power value in a wind speed interval is the normal data, which should be preserved while other clusters are eliminated.

IV. Case Study

A. Data Description

The wind speed and wind power data of 20 wind turbines in a wind farm in China are used in this paper. For each wind turbine, more than 15000 *v-p* data pairs are available, which are sufficient for analysis. The cut-in speed is 3.5 m/s, the cut-out speed is 25 m/s and the rated speed is 12 m/s. The rated capacity of the turbines is 2 MW. In the following, a wind turbine numbered 4 is selected for study, since its outliers present typical features of outliers that discussed above.

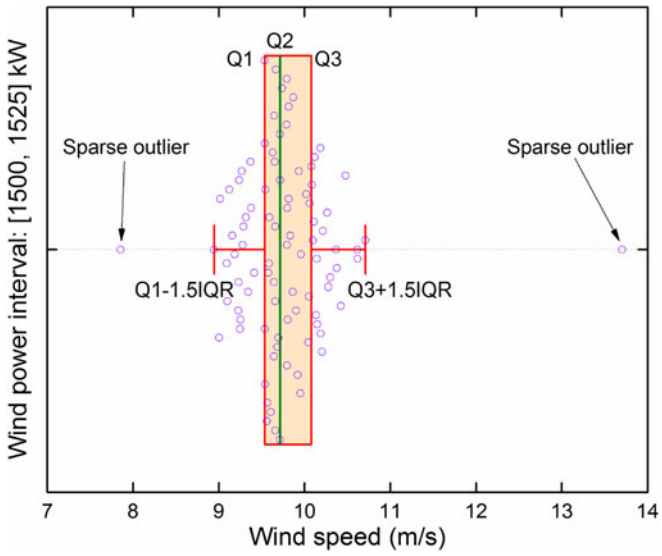


Fig. 5. The elimination of horizontal sparse outliers using quartile method.

The widths of wind speed intervals and wind power intervals are chosen according to a number of trials and suggestion in other literature [25]. They are respectively set as 0.5 m/s and 1.25% of rated power, which is 25 kW. In the following the values of ε will be expressed as percentages of rated power. The two parameters of DBSCAN method, i.e., *MinPts* and ε are set as 5 and 2.5% respectively. The influence of the two parameters on outlier elimination performance will be discussed in more details later.

B. An Example Illustration

1) The Elimination of Sparse Outliers: The elimination of horizontal sparse outliers for wind turbine 4 in the wind power interval [1500, 1525] kW is illustrated in Fig. 5.

The box plot generated by quartile method is given along with the wind speed data points. It can be found that most data points are within the interquartile range. Two points beyond the interquartile range are marked as sparse outliers and will be eliminated.

The scatters after using the preliminary elimination and horizontal and vertical quartile elimination method are shown in Fig. 6. It can be seen that most of the sparse outliers have been eliminated. Most importantly, the boundaries among stacked outliers become clearer. This is very helpful for applying DBSCAN method as it is based on searching for the dense region in the data points.

2) The Elimination of Stacked Outliers: The elimination of stacked outliers in wind speed interval (9.0, 9.5] m/s is illustrated in Fig. 7. In this interval, the wind power data are clustered into three groups. It is clear that the two lower groups should be marked as stacked outliers.

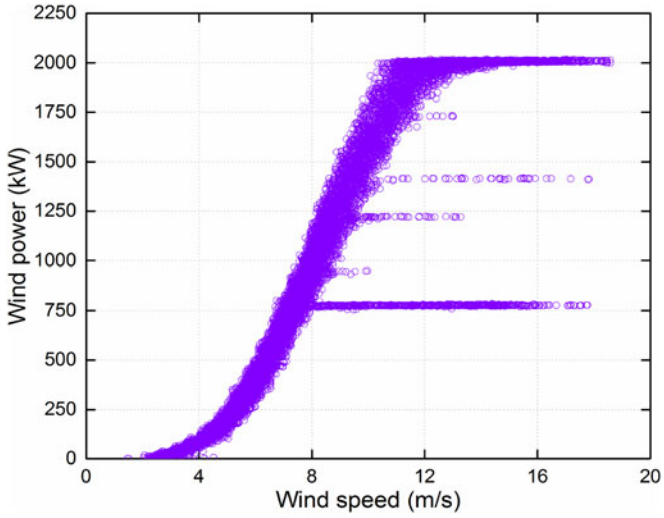


Fig. 6. The *v-p* scatters after preliminary elimination and sparse outlier elimination.

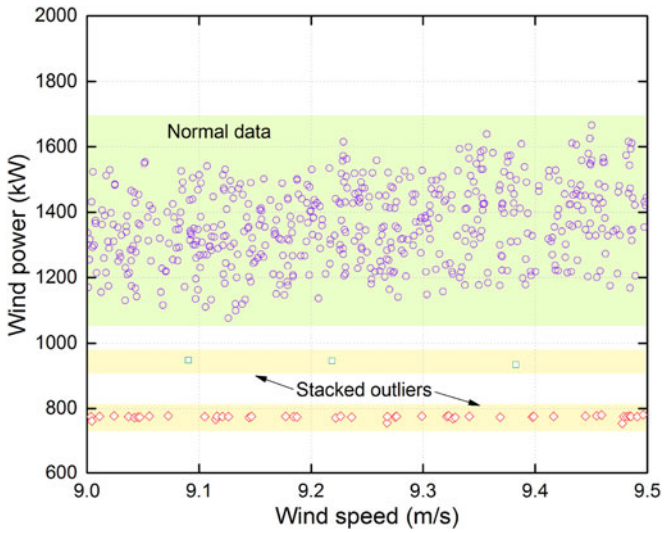


Fig. 7. The elimination of stacked outliers using DBSCAN method.

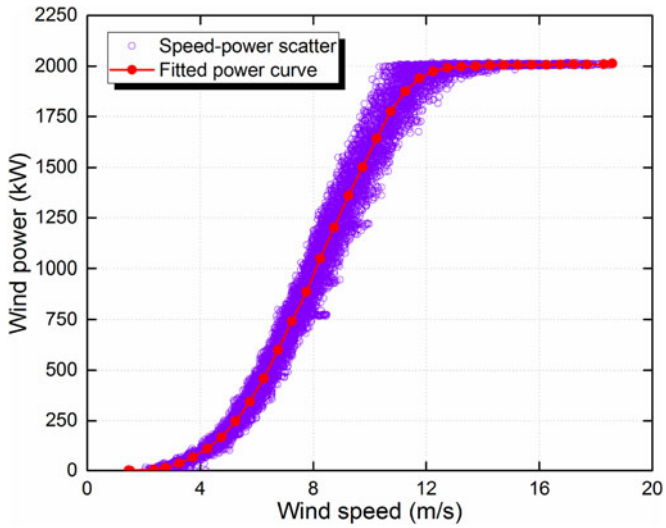


Fig. 8. The *v-p* scatters and fitted power curve after elimination of stacked outliers.

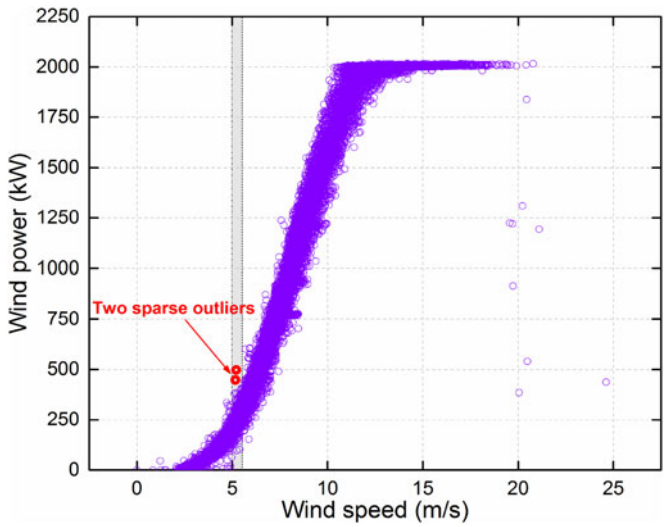


Fig. 9. The elimination result of wind turbine 4 without using quartile method.

The *v-p* scatters and the fitted power curve by "bin" method are shown in Fig. 8. The stacked outliers have been well eliminated and the central dense data region around the power curve has been clearly outlined.

To observe the impact of quartile method on the elimination performance, the outlier elimination result by using only DBSCAN without preprocessing using quartile method is depicted in Fig. 9. The scatters seem very similar to that in Fig. 8. However, the result obtained by using only DBSCAN is very sensitive to the parameters. In Fig. 9, two sparse points are especially marked and indicated. On one hand, with smaller values of *MinPts* (e.g., 2) and ε, the two points will be likely to form a cluster. Then they will be treated as normal data and be preserved because they are in the topmost cluster in the grey colored wind speed interval, whereas other data below the two marked points will be seen as outliers and eliminated. On the other hand, even if *MinPts* keeps as 5, once there are at least 5 close points in original dataset at the topmost of the interval, they will also form a cluster and the normal data below them will be wrongly eliminated. It can be deduced that if there are more than *MinPts* sparse points but close enough to form a cluster at the topmost of a wind speed interval, then normal data will be eliminated while outliers will be preserved by using only DBSCAN.

The elimination results of wind turbine 7 is additionally presented, which can better illustrate why quartile method is necessary before using DBSCAN. With the same parameters, *MinPts* $= 5$, $\varepsilon = 2.5\%$, the elimination results without using quartile method before DBSCAN are presented in the Fig. 10. It shows that, without using quartile method, the normal data in the wind speed interval are eliminated, due to the topmost cluster formed by 6 sparse outliers. This situation can be improved by using quartile method to eliminate sparse outliers before DBSCAN, as given in Fig. 11. Thus, quartile method is very helpful for DBSCAN to eliminate the stacked outliers and makes the elimination procedure more robust and more insensitive to parameters.

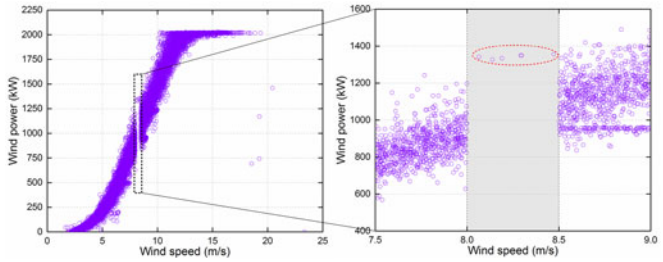


Fig. 10. The elimination result of wind turbine 7 without using quartile method before DBSCAN.

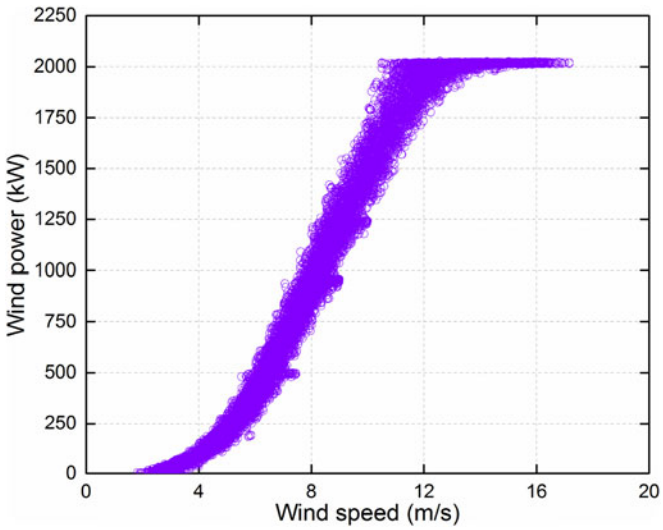


Fig. 11. The elimination result of wind turbine 7 using both quartile method and DBSCAN.

C. Performance Index for Methods Evaluation

Two metrics are used to measure the performance of proposed outlier elimination methods: power curve modeling error e_{M} and elimination rate γ. e_{M} is expressed as normalized root mean square error (NRMSE):

$$e_{\mathrm{M}} = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\hat{p}_i - p_i\right)^2}}{P_{\mathrm{r}}} \times 100\% \tag{7}$$

where N is the number of data points used for evaluation, p_i is the wind power measurement at a given wind speed, $\hat{p}_i$ is the calculated wind power by using "bin" curve modeling at the same wind speed as that of p_i. As the modeled wind power curve using "bin" method is discrete, the cubic spline interpolation is applied to obtain a continuous power curve.

The elimination rate is defined as the ratio between the number of eliminated outliers and the number of data points before elimination, that is

$$\gamma = \frac{N_{\mathrm{b}} - N_{\mathrm{a}}}{N_{\mathrm{b}}} \times 100\% \tag{8}$$

where N_{a} is the number of data points after elimination, N_{b} is the number of data points before elimination. For a well performed outlier elimination method, both e_{M} of the power curve modeling and the elimination rate γ should be as small as possible.

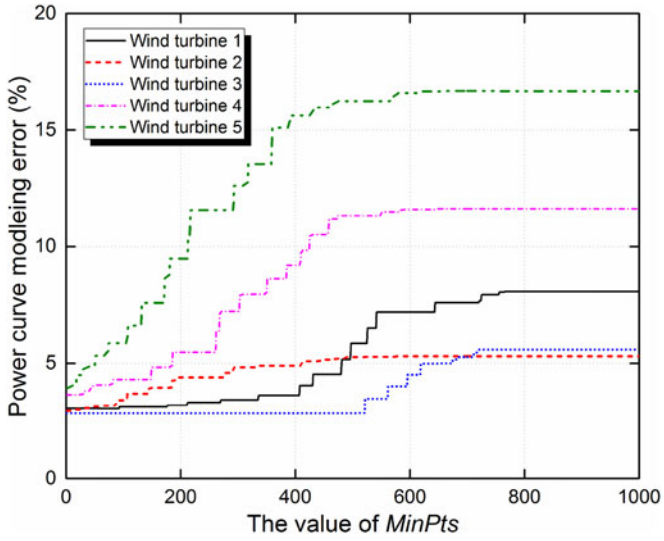


Fig. 12. The trend of power curve modeling error with varying value of *MinPts*.

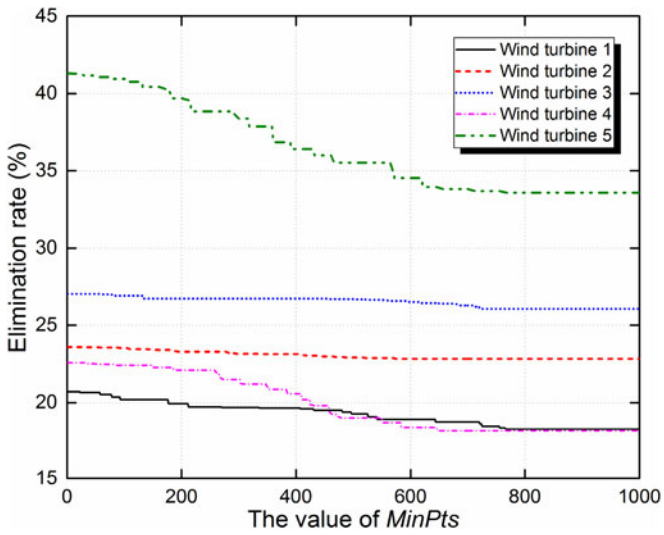


Fig. 13. The trend of elimination rate with varying value of *MinPts*.

D. The Parameters of DBSCAN

The most important parameters for DBSCAN method are *MinPts* and ε. The choices of these two parameters are dependent on many factors, such as the number of the outliers, the features of the outliers, the distances between stacked outliers and normal data, etc. For example, if the value of ε is too small, the data will be over-eliminated, i.e., too many normal data points will be removed. If the value of ε is too large, the DBSCAN may lose its function and will not eliminate any stacked outliers.

The trends of power curve modeling errors and outlier elimination rates for five wind turbines (Turbine 1 to 5) in terms of the value of *MinPts* are depicted in Figs. 12 and 13. To observe these relationships, the ε is fixed as 2.5%.

In Fig. 12, for all the five wind turbines, the power curve modeling errors tend to increase with the increasing *MinPts*. After a certain value of *MinPts*, the modeling error no longer changes even if *MinPts* continue increasing. In Fig. 13, the elimination

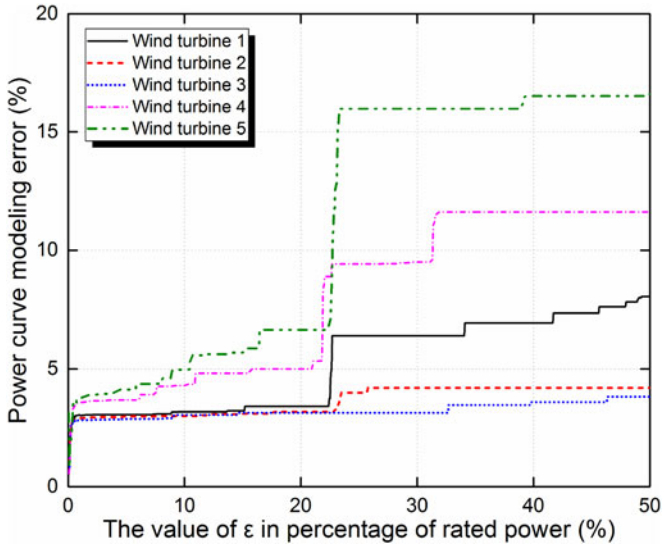


Fig. 14. The trend of power curve modeling error with varying value of ε.

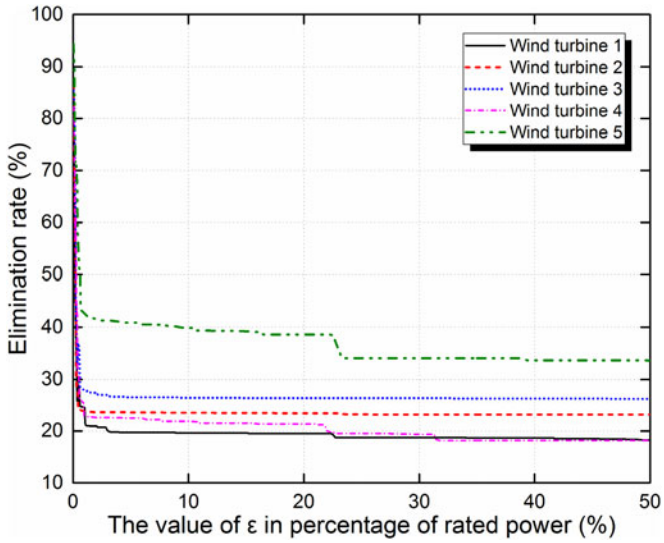


Fig. 15. The trend of elimination rate with varying value of ε.

rate presents downward trend with increasing *MinPts* and then keeps unchanged for large values of *MinPts*. This corresponds to the trend of modeling error. When *MinPts* is too large, stacked outliers cannot be eliminated by DBSCAN, causing low elimination rate and high modeling error level. It should be noted that there is a plateau with different length at the starting of each modeling error curve. Taking wind turbine 3 as example, its modeling error is low and keeps almost constant when *MinPts* varies from 1 to more than 500. The length of plateau is about 100 for turbine 1 and turbine 2, about 50 for turbine 4 and about 20 for turbine 5.It is concluded that a range of small values are available for setting *MinPts* and all the *MinPts* in this range can ensure a stable performance of DBSCAN for most of wind turbines, though the elimination rates at these values are high. In this paper, the *MinPts* for all wind turbines are set as 5.

The trends of power curve modeling errors and elimination rates in terms of the value of ε are depicted in Figs. 14 and 15.

It seems that the relationship between power curve modeling error and the value of ε is more complex. Notably, the modeling errors are very small for $\varepsilon < 0.25\%$. However, ε cannot be set as so small values because the corresponding elimination rate is extremely high. In fact, the elimination rates for all the five wind turbines are higher than 80% at $\varepsilon = 0.05\%$. This means that data is over-eliminated by DBSCAN and no sufficient data is left for modeling, thus this result is unreliable. From Fig. 14, it can be seen that the modeling error curves also have plateaus after $\varepsilon > 0.25\%$. It is reasonable to choose ε among $1\% < \varepsilon < 10\%$ to ensure the stable performance of DBSCAN. In this paper, ε is set as 2.5%.

E. The Results of Proposed Outlier Elimination Methods

The values of e_{M} before and after outlier elimination for 20 wind turbines are given in Table I. As shown in Table I, the power curve modeling errors after outlier elimination for all the 20 wind turbines decrease significantly in comparison with that before outlier elimination. This indicates that the outlier elimination method proposed in this paper is effective for improving the data quality and thus for improving the power curve modeling. The method shows good generality and can be applied to different wind turbines without elaborately tuning parameters. All of the 20 wind turbines can achieve good performance by sharing the same DBSCAN parameters.

The elimination rates of 20 wind turbines are also counted in Table I. Though the method performs well for eliminating both sparse and stacked outliers, it seems that a large proportion of original data has been discarded. For most of the 20 wind turbines, more than 20% of their data is eliminated. Especially for wind turbine 5, the elimination rate even reaches up to 41.31%. The high elimination rates will badly damage the integrality of the data and also has negative impact on the reutilization of the data. The missing data caused by outlier elimination can be recovered by temporal interpolation methods (e.g., time series forecasting, cubic spline method) and spatial interpolation methods [33]–[35]. In particular, the newly built power curve model after outlier elimination can also be used to recover the eliminated data by taking wind speed as inputs [25].

Additionally, the *v-p* scatters of the wind farm (rated capacity is 40 MW) with the 20 wind turbines also present a large number of outliers, as shown in Fig. 16. The fitted power curve seriously deviates from the concentrated data region.

Then, the proposed outlier elimination approach is applied to the wind farm's scatters, for which $MinPts = 5$ and $\varepsilon = 2.5\%$ of wind farm's rated capacity (40 MW). The widths of wind speed intervals and wind power intervals are set as 0.5 m/s and 500 kW, respectively. The scatters after outlier elimination and the fitted power curve are shown in Fig. 17. It can be seen that the outliers have been effectively eliminated and the power curve is well fitted. The elimination rate is 17.88%. The power curve modeling error before and after outlier elimination is 14.17% and 5.1% respectively. It indicates that the accuracy of wind farm power curve modeling can also be significantly improved by eliminating the outliers in the scatters.

TABLE I
THE POWER CURVE MODELING ERRORS, ELIMINATION RATES AND COMPUTATION TIME OF 20 WIND TURBINES

Turbine Number	e_M (%)		γ (%)	Computation time (s)	Turbine Number	e_M (%)		γ (%)	Computation time (s)
	Before elimination	After elimination				Before elimination	After elimination		
1	16.94	3.05	20.71	0.653	**11**	13.25	4.15	16.15	0.977
2	19.09	2.95	23.60	0.835	**12**	13.78	2.84	25.35	0.985
3	18.93	2.84	27.02	0.723	**13**	13.55	3.43	16.02	1.023
4	21.37	3.65	22.57	0.808	**14**	15.99	5.98	21.94	0.733
5	23.67	3.92	41.31	0.610	**15**	16.08	4.84	21.92	0.720
6	19.55	3.94	27.29	0.802	**16**	18.25	3.69	30.07	0.862
7	17.22	3.69	25.75	0.908	**17**	19.25	3.71	29.99	0.824
8	19.35	3.60	19.08	0.764	**18**	15.10	3.82	21.18	0.852
9	19.67	3.47	20.47	0.795	**19**	16.47	4.40	18.73	0.748
10	21.33	4.26	27.27	0.703	**20**	18.63	3.74	33.65	0.716

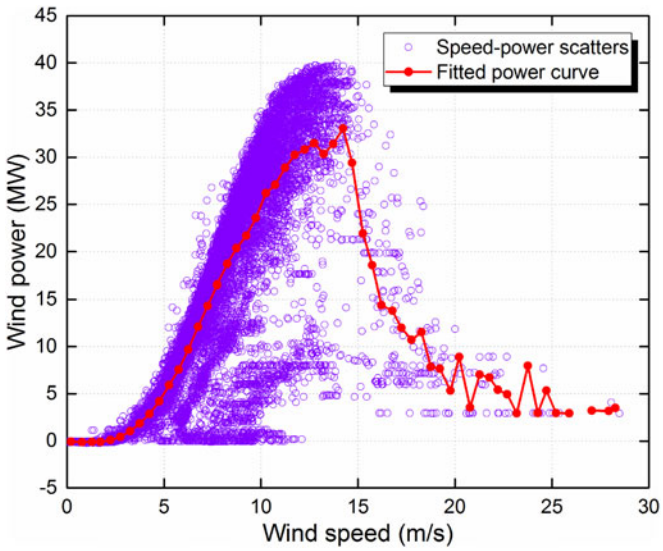


Fig. 16. The *v-p* scatters of wind farm and the fitted power curve with outliers.

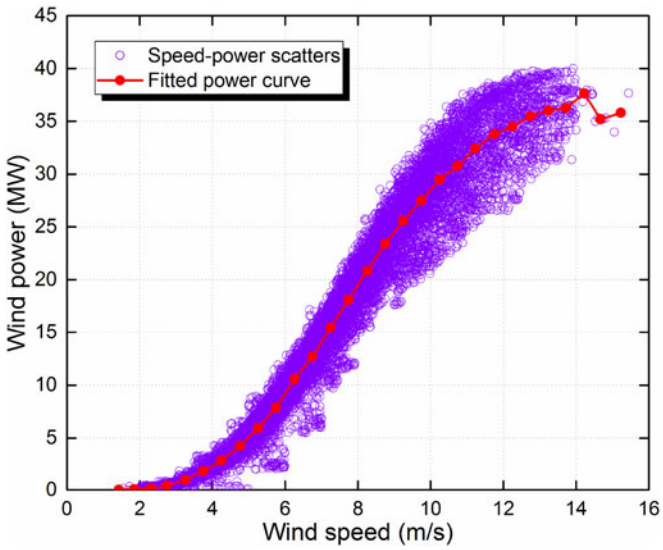


Fig. 17. The *v-p* scatters of wind farm and the fitted power curve after outlier elimination.

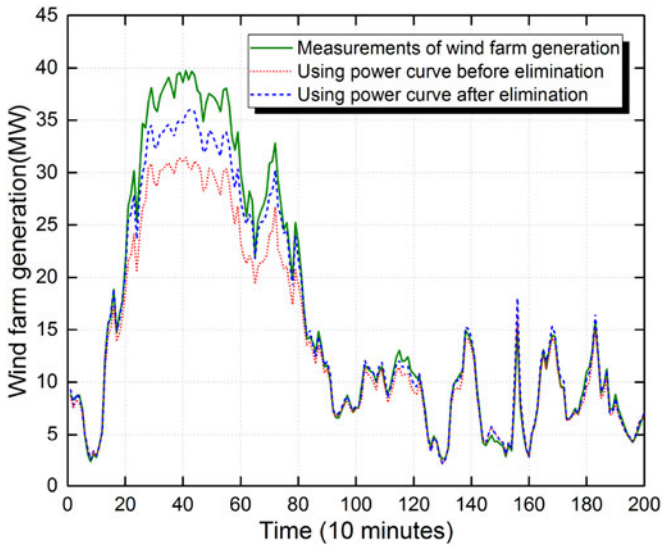


Fig. 18. The predicted wind power output of the wind farm using power curves before and after outlier elimination.

The cubic spline interpolation method is applied to obtain continues power curves. Then the time series of wind speed over a period is input to the power curves before and after outlier elimination to predict the wind power output of the wind farm. The wind power outputs and their corresponding measurements are compared in Fig. 18. It can be seen that the predicted wind power by using power curve after outlier elimination is more accurate than that before outlier elimination. The difference is more obvious at high levels of wind power generation, where wind curtailment always occurs.

The efficiency of an algorithm is a very important factor that should be considered. The proposed method is implemented in Matlab 2010 b (64 bit) on a PC with 2.9 GHz Intel Dual Core CPU and 4 GB RAM. The algorithm for each wind turbine is executed 5 times and the average of 5 computation times for the 20 wind turbines are given in Table I. It shows that the elimination of outliers for 19 wind turbines can be finished within 1 second though Turbine 13 takes 1.023 seconds, which reveals the high efficiency of the algorithm.

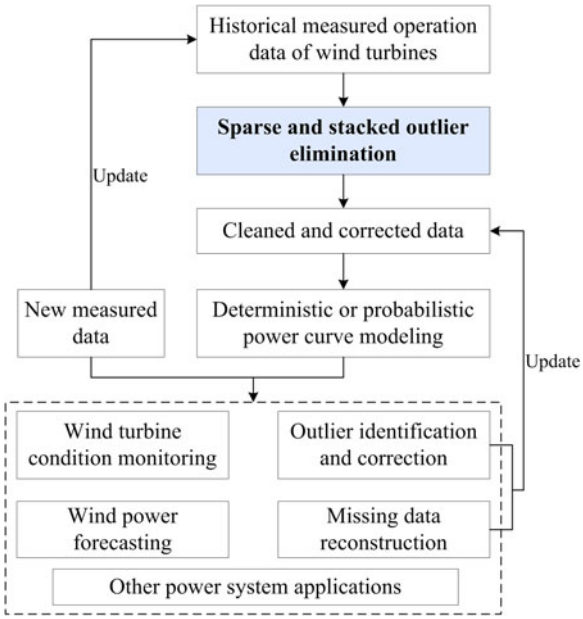


Fig. 19. A chain of potential online applications involving outlier elimination.

F. The Online Applications

A chain of potential online applications involving outlier elimination is illustrated in Fig. 19.

Cleaned and corrected data can be obtained by implementing sparse and stacked outlier elimination based on historical measured operation data of wind turbines. With the cleaned dataset, a deterministic or probabilistic power curve model can be built using existing or more advanced methods. When a new measured operation data pairs of wind speed and wind power is available, it is input to the power curve model and checked whether it is an abnormal data. At the same time, the wind turbine condition can also be assessed by data checking. The abnormal data will be further corrected to the value derived from power curve model. Sometimes, the missing wind power data can be fulfilled by the value calculated from power curve model with the input of the corresponding wind speed. The new measured data will be added to the historical measured operation database. With increasing amount of updated historical data, the sparse and stacked outlier elimination method can be more reliable when it is implemented in next loop. The corrected data from outlier elimination, correction and missing data reconstruction will be added to cleaned database to get more accurate power curve model. The power curve can be extensively applied to wind power forecasting and other power system applications as needed.

V. Conclusion

In this paper, the characteristics of outliers in *v-p* scatters are analyzed. The stacked outliers caused by wind curtailment can deform the relationship between wind speed and power output and lead to low accuracy of power curve modeling. To this end, a combined approach is proposed to eliminate the outliers in the scatters, including sparse outliers and stacked outliers. The quartile method is used twice to eliminate the sparse outliers. This process can make the boundaries among stacked outliers clearer, which is very helpful for eliminating stacked outliers. Then, the DBSCAN method is applied to eliminate stacked outliers. The combined approach is totally data-driven. There is no need to impose manufacture's specifications or statistical assumptions on the approach.

The power curve modeling error and the outlier elimination rate are introduced to quantify the performance of outlier elimination. Results show that the combined outlier elimination approach is effective and efficient for both wind turbines and wind farm. What is also worth noticing is that the performance of the DBSCAN is insensitive to its parameters, revealing that it can be directly used in different cases without carefully tuning the parameters.

However, the outlier rejection rate is high, for which missing data reconstruction methods need to be studied and applied to ensure the integrity of dataset. It should be noted that the DBSCAN parameters for all the wind speed intervals of a wind turbine are sharing a same value. This may not be the best option since there are differences among the outlier characteristics of different wind speed intervals. Thus it is possible to set different parameters for each individual interval according to their data features, from which better results could be obtained.

References

[1] Z. Li, L. Ye, Y. Zhao, X. Song, J. Teng, and J. Jin, "Short-term wind power prediction based on extreme learning machine with error correction," *Protection Control Modern Power Syst.*, vol. 1, no. 1, pp. 1–8, Jun. 2016.

[2] Y. Zhao, L. Ye, Z. Li, X. Song, Y. Lang, and J. Su, "A novel bidirectional mechanism based on time series model for wind power forecasting," *Appl. Energy*, vol. 177, pp. 793–803, Sep. 2016.

[3] M. Milligan, E. Ela, D. Lew, D. Corbus, Y. H. Wan, and B. M. Hodge, "Assessment of simulated wind data requirements for wind integration studies," *IEEE Trans. Sustain. Energy*, vol. 3, no. 4, pp. 620–626, Oct. 2012.

[4] D. Villanueva, A. Feijoo, and J. L. Pazos, "Simulation of correlated wind speed data for economic dispatch evaluation," *IEEE Trans. Sustain. Energy*, vol. 3, no. 1, pp. 142–149, Jan. 2012.

[5] M. Lydia, S. S. Kumar, A. I. Selvakumar, and G. E. P. Kumar, "A comprehensive review on wind turbine power curve modeling techniques," *Renewable Sustain. Energy Rev.*, vol. 30, pp. 452–460, Feb. 2014.

[6] M. Lydia, A. I. Selvakumar, S. S. Kumar, and G. E. P. Kumar, "Advanced algorithms for wind turbine power curve modeling," *IEEE Trans. Sustain. Energy*, vol. 4, no. 3, pp. 827–835, Jul. 2013.

[7] S. Shokrzadeh, M. J. Jozani, and E. Bibeau, "Wind turbine power curve modeling using advanced parametric and nonparametric methods," *IEEE Trans. Sustain. Energy*, vol. 5, no. 4, pp. 1262–1269, Oct. 2014.

[8] S. Gill, B. Stephen, and S. Galloway, "Wind turbine condition assessment through power curve copula modeling," *IEEE Trans. Sustain. Energy*, vol. 3, no. 1, pp. 94–101, Jan. 2012.

[9] M. Schlechtingen, I. F. Santos, and S. Achiche, "Using data-mining approaches for wind turbine power curve monitoring: A comparative study," *IEEE Trans. Sustain. Energy*, vol. 4, no. 3, pp. 671–679, Jul. 2013.

[10] A. Kusiak, H. Y. Zheng, and Z. Song, "Models for monitoring wind farm power," *Renewable Energy*, vol. 34, pp. 583–590, Mar. 2009.

[11] A. Kusiak and A. Verma, "Monitoring wind farms with performance curves," *IEEE Trans. Sustain. Energy*, vol. 4, no. 1, pp. 192–199, Jan. 2013.

[12] L. Ye, Y. Zhao, C. Zeng, and C. Zhang, "Short-term wind power prediction based on spatial model," *Renewable Energy*, vol. 101, pp. 1067–1074, Feb. 2017.

[13] M. Xu, P. Pinson, Z. Lu, Y. Qiao, and Y. Min, "Adaptive robust polynomial regression for power curve modeling with application to wind power forecasting," *Wind Energy*, vol. 19, pp. 2321–2336, Dec. 2016.

[14] S. Sarkar and V. Ajjarapu, "MW resource assessment model for a hybrid energy conversion system with wind and solar resources," *IEEE Trans. Sustain. Energy*, vol. 2, no. 4, pp. 383–391, Oct. 2011.

[15] D. Dhungana and R. Karki, "Data constrained adequacy assessment for wind resource planning," *IEEE Trans. Sustain. Energy*, vol. 6, no. 1, pp. 219–227, Jan. 2015.
[16] S. H. Jangamshetti and V. G. Rau, "Normalized power curves as a tool for identification of optimum wind turbine generator parameters," *IEEE Trans. Energy Convers.*, vol. 16, no. 3, pp. 283–288, Sep. 2001.
[17] B. Hu, Y. Li, H. Yang, and H. Wang, "Wind speed model based on kernel density estimation and its application in reliability assessment of generating systems," *J. Modern Power Syst. Clean Energy*, vol. 5, pp. 220–227, 2017.
[18] S. Wang, X. Zhang, L. Ge, and L. Wu, "2-D wind speed statistical model for reliability assessment of microgrid," *IEEE Trans. Sustain. Energy*, vol. 7, no. 3, pp. 1159–1169, Jul. 2016.
[19] N. Yampikulsakul, E. Byon, S. Huang, S. Sheng, and M. You, "Condition monitoring of wind power system with nonparametric regression analysis," *IEEE Trans. Energy Convers.*, vol. 29, no. 2, pp. 288–299, Jun. 2014.
[20] J. Y. Park, J. K. Lee, K. Y. Oh, and J. S. Lee, "Development of a novel power curve monitoring method for wind turbines and its field tests," *IEEE Trans. Energy Convers.*, vol. 29, no. 1, pp. 119–128, Mar. 2014.
[21] X. Ye, Z. Lu, Y. Qiao, Y. Min, and M. O'Malley, "Identification and correction of outliers in wind farm time series power data," *IEEE Trans. Power Syst.*, vol. 31, no. 6, pp. 4197–4205, Nov. 2016.
[22] Y. Wan, E. Ela, and K. Orwig, "Development of an equivalent wind plant power curve," in *Proc. WindPower*, 2010, pp. 1–20.
[23] Y. Wang, D. G. Infield, B. Stephen, and S. J. Galloway, "Copula-based model for wind turbine power curve outlier rejection," *Wind Energy*, vol. 17, pp. 1677–1688, Nov. 2014.
[24] E. Mangalova and E. Agafonov, "Wind power forecasting using the k-nearest neighbors algorithm," *Int. J. Forecasting*, vol. 30, pp. 402–406, Apr.–Jun. 2014.
[25] D. Villanueva and A. Feijóo, "Normal-based model for true power curves of wind turbines," *IEEE Trans. Sustain. Energy*, vol. 7, no. 3, pp. 1005–1011, Jul. 2016.
[26] Wind Turbines—Part 12-1: Power Performance Measurements of Electricity Producing Wind Turbines, IEC-International Electrotechnical Commission, IEC-61400-12, 2005.
[27] Y. Gu and L. Xie, "Fast sensitivity analysis approach to assessing congestion induced wind curtailment," *IEEE Trans. Power Syst.*, vol. 29, no. 1, pp. 101–110, Jan. 2014.
[28] L. Gan, G. Li, and M. Zhou, "Coordinated planning of large-scale wind farm integration system and transmission network," *CSEE J. Power Energy Syst.*, vol. 2, no. 1, pp. 19–29, Mar. 2016.
[29] D. S. Moore, G. P. Mccabe, and B. A. Craig, *Introduction to the Practice of Statistics*. New York, NY, USA: Freeman, 2012.
[30] R. J. Hyndman and Y. N. Fan, "Sample quantiles in statistical packages," *Amer. Statist.*, vol. 50, pp. 361–365, Nov. 1996.
[31] B. S. Everitt, S. Landau, M. Leese, and D. Stahl, *Cluster Analysis*, 5th ed. Hoboken, NJ, USA:Wiley, 2011.
[32] M. Ester, H. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. 2nd Int. Conf. Knowl. Discovery Data Mining*, 1996, pp. 226–231.
[33] Q. Lin and J. Wang, "Vertically correlated echelon model for the interpolation of missing wind speed data," *IEEE Trans. Sustain. Energy*, vol. 5, no. 3, pp. 804–812, Jul. 2014.
[34] Y. Zhang, S. J. Kim, and G. B. Giannakis, "Short-term wind power forecasting using nonnegative sparse coding," in *Proc. 2015 49th Annu. Conf. Inf. Sci. Syst.*, Baltimore, MD, USA, 2015, pp. 1–5.
[35] Z. Yang, Y. Liu, and C. Li, "Interpolation of missing wind data based on ANFIS," *Renewable Energy*, vol. 36, pp. 993–998, Mar. 2011.

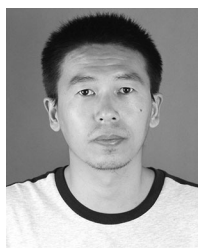
Yongning Zhao (S'17) received the B.Sc. degree in electrical engineering from China Agricultural University, Beijing, China, in 2012. He is currently working toward the Ph.D. degree with the Department of Electric Power Systems, China Agricultural University, China. He is currently a visiting student with the Center for Electric Power & Energy, Technical University of Denmark, Lyngby, Denmark. His research interests include the areas of power system operation and control, analysis of the spatial-temporal characteristics of wind power, wind power forecasting and integration.

Lin Ye (M'94–SM'06) received the B.Sc. degree from WuHan University, Wuhan, China, in 1992 and the Ph.D. degree from the Institute of Electrical Engineering (IEE), Chinese Academy of Sciences (CAS), Beijing, China, in 2000, both in electrical engineering.

He has been pursuing research at ForschungsZentrum Karlsruhe (now merged with University of Karlsruhe(TH) to form Karlsruhe Institute for Technology, KIT) as a Research Fellow of Alexander von Humboldt Stiftung/Foundation (AvH) of Germany from 2000 to 2002. He joined the Interdisciplinary Research Center, Department of Engineering/Cavendish Laboratory, the University of Cambridge, U.K., as a Research Fellow in 2004. At Cambridge Laboratory, he had been involved in developing a novel resistive type of superconducting fault current limiter prototype for electrical marine propulsion, which was funded by Rolls-Royce Plc and the Department of Trade & Industry of the United Kingdom. He is currently a Full Professor in electrical power engineering at the Department of Electric Power Systems, China Agricultural University, Beijing, China. His research interests include electric power system analysis & control, power grids modeling & simulations, renewable energy generation & system integration, and wind/solar power forecasting. He holds memberships in IEEE (USA), and European EMTP-ATP Users Group (EEUG) as well as a senior member of Wolfson College at Cambridge University. He is also a core member of the IEEE Task Force on Sustainable Energy Systems for Developing Communities and an international expert in CIGRE SC B5 working group. He received the "Hua Wei" Award from Chinese Academy of Sciences in 1999, and the Research Young Investigator Award from Fok Ying Tung Education Foundation, Ministry of Education, China in 2003. He was an awardee of the Program for New Century Excellent Talents in China Universities in 2008.

Weisheng Wang (M'09–SM'15) received the Doctor degree in electrical engineering at Xi'an Jiaotong University, Xi'an, China, in 1996. Then, he joined China Electric Power Research Institute (CEPRI), Beijing, China, in January 1997. Currently, he is a Professor and the Director of Renewable Energy Department of CEPRI. His main interests include research and consulting in the field of renewable energy generation and its grid integration.

Huadong Sun (SM'15) received the B.E. and M.S. degrees in electrical engineering from Shandong University, Jinan, China, in 1999 and 2001, respectively, and the Ph.D. degree from China Electric Power Research Institute (CEPRI), Beijing, China, in 2005. He is currently a Professor at the CEPRI. His current research interest include power system security analysis and control.



Yuntao Ju (M'13) received the B.Sc. degree in mechanical engineering in 2008 and the Ph.D. degree in electrical engineering in 2013, all from Tsinghua University, Beijing, China. He has been a visiting scholar to the University of Toronto for a year. In 2015, he joined the china electric power research institute as a Research Fellow. He is currently an Associate Professor at the College of Information and Electrical Engineering, China Agricultural University, Beijing, China. His research interests include hybrid energy system modeling, high-speed dynamic simulation, large-scale system parameter identification, state estimation, and uncertainty optimization.

Yong Tang (M'09–SM'15) received the M.E. and Ph.D. degrees from China Electric Power Research Institute (CEPRI), Beijing, China, in 1985 and 2002, respectively. He is currently a Professor and a Ph.D. Supervisor with CEPRI. His research interests include the area of power system simulation and analysis, voltage stability and control, load modeling and simulation.