energies

MDPI

Article

An Improved Data-Efficiency Algorithm Based on Combining Isolation Forest and Mean Shift for Anomaly Data Filtering in Wind Power Curve

Wei Wang, Shiyou Yang * and Yankun Yang

College of Electrical Engineering, Zhejiang University, Hangzhou 310027, China; d77mvp@zju.edu.cn (W.W.); 22110081@zju.edu.cn (Y.Y.)

* Correspondence: eesyyang@zju.edu.cn

Abstract: A wind turbine working in a harsh environment is prone to generate abnormal data. An efficient algorithm based on the combination of an Isolation Forest (I-Forest) and a mean-shift algorithm is proposed for data cleaning in wind power curves. The I-Forest is used for detecting the local anomalies in each power and wind speed interval after data preprocessing. The contamination of I-Forest can be flexibly adjusted according to the data distribution of the wind turbine data. The remaining stacked data is eliminated by the mean-shift algorithm. To verify the filtering performance of the proposed combined method, five different algorithms, including the quartile and *k*-means (QK), the quartile and density-based spatial clustering (QD), the mathematical morphology operation (MMO), the fast data cleaning algorithm (FA), and the proposed one, are applied to the wind power curves of a prototype wind farm for comparisons. The numerical results have positively confirmed the reliability of the universal framework provided by the proposed algorithm.

Keywords: abnormal data; I-forest; mean-shift; wind power curve; wind turbine

Citation: Wang, W.; Yang, S.; Yang, Y. An Improved Data-Efficiency Algorithm Based on Combining Isolation Forest and Mean Shift for Anomaly Data Filtering in Wind Power Curve. *Energies* **2022**, *15*, 4918. https://doi.org/10.3390/en15134918

Academic Editor: Konstantin Suslov

Received: 13 June 2022
Accepted: 2 July 2022
Published: 5 July 2022

1. Introduction

Wind energy is widely utilized in modern electricity generation due to its environment-friendly and renewable energy nature. Compared with thermal power, wind power is often characterized by strong volatility and randomness, which can affect the stable operation of a power system. It is thus essential to predicate wind power to guarantee a safe operation of the power system. The wind power data recorded by the Supervisory Control and Data Acquisition System (SCADA) is an input variable for very short-term power prediction [1–6]. Due to the fact that wind turbines are generally located in an open-air field and work in a chaotic environment, they are easily influenced by weather conditions. Accordingly, it is necessary to remove irrational data from the SCADA before power forecasting.

There exist numerous studies in wind power anomaly data filtering. These methods can be divided into three categories. The first category mainly explores the distribution features of abnormal data in order to select an appropriate algorithm for data cleaning [7–13]. Shen et al. proposed a combined algorithm made by changepoint grouping and a quartile algorithm [7]. Changepoint grouping determines the gap between the normal data and the abnormal data in a specific wind speed interval. However, there may exist multiple points in some special intervals, with large variance change rates by using the changepoint grouping, which may eventually lead to a false elimination. Xiang et al. applied the quartile method to filter abnormal data [8]. However, this method is only applicable to wind turbines with few abnormal data. In [9], the quartile method was used to clean scattered data; the double *k*-means clustering was employed to eliminate stacked data. However, the value of parameter *k* is difficult to determine and may result in clustering failure. In [10], the author replaced the *k*-means with the density-based spatial clustering of applications with noise (DBSCAN). DBSCAN does not need to manually formulate the classification

number. However, the two parameters of a DBSCAN are sensitive to the intervals with small derivatives between normal data and abnormal data. Zheng et al. divided the data into six categories and employed the local outlier factor to distinguish normal data and abnormal data [11]. However, when the wind turbine contains a large amount of abnormal data, this method may require a lot of solution time [7], and the accuracy may not be guaranteed. Sahra Khazaei et al. utilized an automatic clustering method and T^2 statistic to detect the outlier [13]. However, this method is only applicable to wind turbines with a small amount of abnormal data.

The work in the second category is to determine the upper and lower boundaries of the wind power curve. The data falling outside the curve boundary is regarded as abnormal data. This type of method basically involves the probabilistic and quantile model. Hu et al. [14] used an adaptive confidence boundary modeling process of wind power curves to remove the abnormal data. Ye et al. [15] built a wind power curve model via the copula condition function. Guo et al. [16] assumed that every wind power data follows a Gaussian distribution at the collection time and constructs the correct wind power curve by using the Gaussian process. Such assumptions often need to combine other external variables, such as rotor-angular velocity, and it requires abundant normal data to train the ideal wind power curve.

The third state-of-art design technique is image processing for wind power curves [17–20]. Long et al., [17] tuned the wind power curve into a binary image after data preprocessing. Normal data are acquired by the maximum contour of the binary image dealt with by the mathematical morphology operation (MMO). However, when the stacked data are closely distributed with normal data, the opening operation may not perform well. Wang et al., in [18], proposed a fast data cleaning algorithm (FA) to keep the longest continuity pixels for each column and row in the binary curve image. However, when the wind curtailment is particularly serious, the longest pixel sequence may appear at the bottom of the column, which mistakenly eliminates the normal pixel sequence. Su et al. developed a pixel counting method to generate the feature gray image [19,20]. The abnormal data can be eliminated via image thresholding segmentation. This algorithm can obtain a valid result in cases where only a wind curtailment is not serious.

To conclude, although different efforts on removing abnormal data have been made in literature, the technique for eliminating abnormal data in the wind power curve with a high proportion of stacked data is still demanding. In this respect, we propose a novel combined algorithm based on Isolation Forest and mean-shift to detect anomalous data in wind turbines. First, one conducts data preprocessing for wind turbines and divides the power and wind speed into several intervals, respectively. The v–p scatter points in each interval are checked by the Isolation Forest algorithm. This process can eliminate the scattered data and some stacked data. The mean-shift algorithm is followed to clear up the rest of the abnormal data. The reliability of the proposed algorithm is validated by the numerical experiment of four wind turbines on a wind farm.

2. Wind Power Curves

The wind power curve is mainly used to analyze the mapping relationship between wind speed and power. Under ideal conditions, when the wind speed is smaller than the cut-in wind speed or higher than the cut-out wind speed, the wind turbines will not generate electricity. When the wind speed is between the cut-in wind speed and the cut-out wind speed, the output power is a cubic function of the input wind speed. Since the weather condition is full of uncertainty, the actual wind power curve seldom follows this ideal model. Moreover, the abnormal data distribution of a wind turbine is diverse due to various working conditions. For example, the four wind power curves for the No. 3, No. 4, No. 5 and No. 6 turbines from an actual dataset that includes twelve wind turbines, as given in [21], are shown in Figure 1. The data in the four wind turbines cover nearly one year; the data were collected every 10 min. The No. 3 data were recorded from 1 November 2017 to 30 October 2018; No. 4 data, from 1 November 2018 to 17 October 2019; No. 5 data,

from 2 February 2019 to 31 December 2019; and No. 6 data, from 1 November 2017 to 30 October 2018. The cut-in wind speed of each wind turbine was 3 m/s, and the cut-out wind speed was 25 m/s. The rated power was 2000 kW. Considering the actual working conditions of the wind turbine, the normal power range of the wind turbine is from 0 to 1.2 times the rate of power.

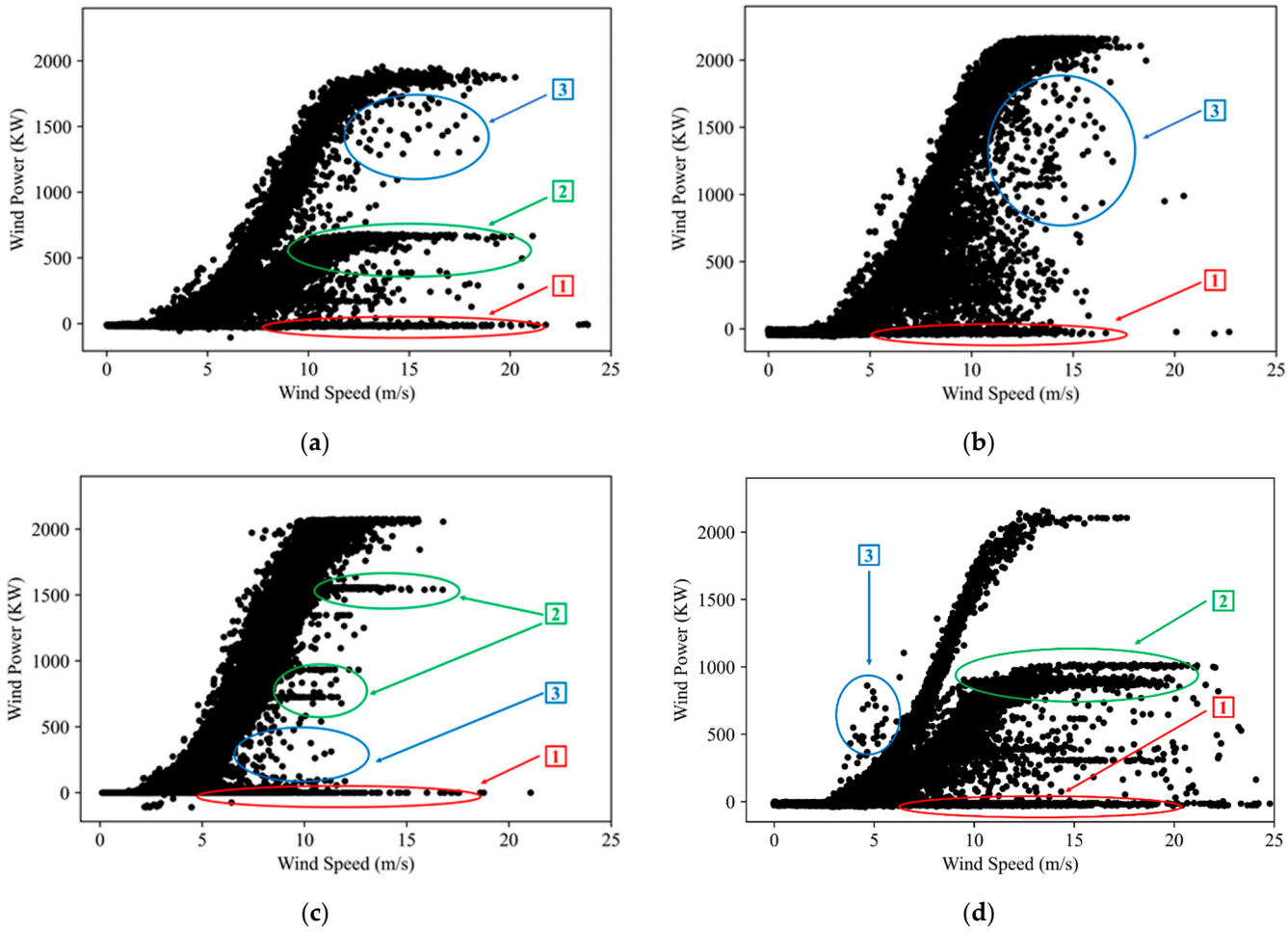

Figure 1. Wind power curves of four wind turbines: (**a**) No. 3, (**b**) No. 4, (**c**) No. 5, (**d**) No. 6.

As introduced in [17], three types of outliers can be found in Figure 1. The No. 1 type (red) is often constituted by negative data. The No. 2 type (green) represents the stacked data, which originates from the dispatching instruction of the power system, and most of the anomalous data belong to this class. The No. 3 type (light blue) is scattered data, caused by random factors [7]. Generally, the stacked data are horizontally distributed on the right side of the normal power curve. However, the shape of stacked data in Figure 1a,d are similar to a compressed normal power curve, thus, the normal data and abnormal data are extremely close in the low-wind speed range. In these cases, the content of abnormal data is much more than that of ordinary wind turbines, such as Figure 1c. It is difficult to completely clean this compressed abnormal data by using the previous algorithms.

3. Combined Isolation Forest and Mean-Shift Algorithm

3.1. Data Preprocessing

In this procedure, when the wind speed is less than the cut-in wind speed or greater than the cut-out wind speed, the wind turbine will not generate power; therefore, in these two cases, the data with power greater than 0 should be filtered. In addition, the negative power data is also unrealistic and should be detected as outliers.

3.2. Isolation Forest Algorithm for Data Cleaning

The Isolation Forest (I-forest) algorithm [22] can quickly detect anomalous data. This method is widely used in a variety of anomaly detection schemes [23,24]. The traditional anomaly data detection, such as the quartile method, is aimed at calculating the area in the feature space. The data will be deemed normal data if they are dropped within this area. I-Forest will not explore such an area to describe the normal data. In I-forest, the anomalies are defined as more likely to be separated [22]. This algorithm assumes that the normal points are usually clusters with a high density, and the abnormal data only account for a small amount, and the attribute of abnormal data is very different from that of the normal data. High-density clusters can be isolated for a limited number of times, while the low-density points can easily be isolated. The mechanism of the I-forest is similar to the random forest. It utilizes a method called isolation tree (iTree) to segment samples.

The single iTree is a random binary tree. Its training process mainly includes the following steps:

1. One selects n points $X = \{x_1, x_2, \ldots x_n\}$ randomly from the training data as a subsample that is set as the root of iTree. Every sample dimension is d.
2. One selects an attribute in d from these data and generates a value p between the maximum and minimum values of the specified attributes in the current root data.
3. This segmentation point p divides the subsample into two parts: The points in the subsample whose attribute value is less than p are placed in the left branch of the current node, while the points whose attribute value is greater than or equal to p are placed in the right branch of the current node.
4. Repeat steps 2 and 3 recursively on the left and right branches of this node. Continue to create new nodes until the leaf node has only one data (it is unable to cut data), or the height of the tree reaches the set limit.

Because the selected feature and segmentation point values are random, it is not advisable to judge that one specific point is the outlier, even if this point in an iTree is abnormal. In order to avoid the influence of contingency, one can randomly re-sample the dataset to build multiple iTrees via the training process. Many iTrees constitute the forest. After the forest is constructed, each xi point in the sample will gain experience through each iTree. The path length of each xi point in iTree is recorded, and the expectation of all path lengths can be calculated. The abnormal score of each point is defined by:

$$s(x, n) = 2^{-\frac{E(h(x))}{c(n)}}. \tag{1}$$

where $h(x)$ represents the path length of x_i passing through on each iTree, and $E(h(x))$ is the expectation of this series of $h(x)$. In addition, n is the number of each subsample. $c(n)$ represents the average path length of the iTree, and is calculated from:

$$c(n) = 2H(n-1) - \frac{2(n-1)}{n}. \tag{2}$$

where $H(i)$ is a harmonic function, estimated by

$$H(i) = \ln(i) + 0.5772156649. \tag{3}$$

The abnormal score $v_{\max}$ of x_i is between 0 and 1. When the score is closer to 1, the path length of x_i in each tree is shorter, and this sample is detected as anomalous data. When this score is close to 0, x_i is identified as normal data. If the scores of most data are close to 0.5, the entire dataset does not contain any abnormal data obviously.

In order to obtain better calculation results, each iTree usually needs to tune some parameters. According to [19], the numbers of iTree and the subsample should be constrained. When the number of iTrees reaches 100 or more, the average path length of each sample point has converged. Therefore, the number of iTrees is generally set to 100. The number

of subsamples is generally set to 256, which can provide sufficient accuracy for anomaly detection and reduce the high computational overhead. Furthermore, I-forests focus on these points with shorter paths, thus, the height of each iTree will be commonly limited to 8.

I-forest has been applied in wind power anomaly data filtering [23], by choosing I-forest to clean the wind power curve in global detection. Fen et al. [23] set the number of iTrees and subsamples to be the same as those in [21]. The rate of outliers in wind turbines is assumed to be 20%. Taking the parameters set in [23] as an example, the results using the Isolation Forest algorithm to detect the global abnormal data of wind turbines, as shown in Figure 1a, are given in Figure 2. It is not advisable to use I-forest to detect abnormal data in wind turbines directly. When the proportion of the abnormal data is high, the effect of the I-forest algorithm will deteriorate sharply.

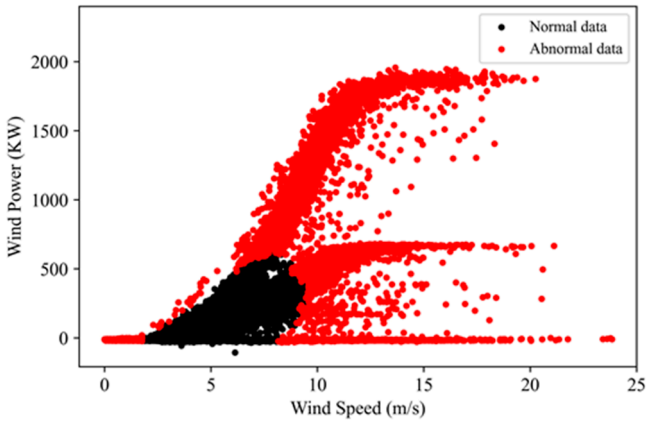

Figure 2. Abnormal data processed by I-Forest in global detection.

Because I-forest is not suitable for global anomaly detection, some improvements must be made. First, these abnormal data, represented by No. 1, can be filtered via data preprocessing. Inspired by [10], one divides the power interval by 25 kW, and the remaining data can be divided into several intervals. One then uses the I-forest to eliminate the abnormal data in every interval. In this process, the contamination ratio in every interval can be set according to the distribution of data, and the number of subsamples in each interval will be determined by the smaller value of sample data and 256 in the interval. Since the I-Forest algorithm is effective for local anomaly data detection, it is appropriate for eliminating scattered data and part of the stacked data which is far from normal data. After this processing, if there are no scattered data around the curve, the mean-shift algorithm is used to eliminate the scattered data, otherwise, one also divides the wind speed range at an interval of 0.5 m/s similarly and processes the outlier in each interval by the I-forest. Since the last operation removed most of the decentralized data, the proportion of scattered abnormal data in each interval has been greatly reduced or has even disappeared. The contamination ratio is supposed to be set into smaller values (e.g., half-percentage) than the last setting, otherwise too much normal data, located in the curve boundary, will also be removed. Finally, one obtains the processed wind power curve, as shown in Figure 3. It is obvious that the anomaly data has been greatly reduced from the wind power curve.

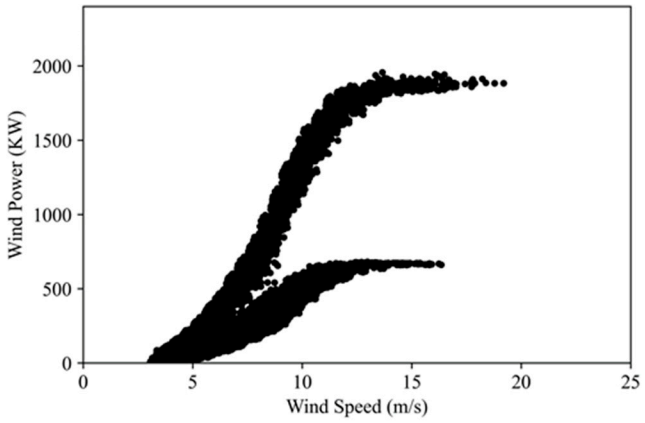

Figure 3. Abnormal data processed by I-Forest in interval detecting.

3.3. Mean-Shift Algorithm for Data Cleaning

After data cleaning processing via I-forests, the main cluster of the remaining abnormal data is the stacked data. Some cluster algorithms were employed in stacked data filtering [9,10]. In order to improve the applicability of these algorithms, Ref [9] utilized the double *k*-means to preserve the data in which the value of its center class was the biggest, and *k* was set to 8. The rational setting of *k* needs to be verified by a large number of wind turbines. In addition, it is difficult to determine the threshold of clustering. Therefore, some cluster algorithms that do not need to specify the number of clusters artificially, for example, the DBSCAN algorithm, are introduced. The distribution of the stacked data and normal data in each interval was analyzed in Ref [10]. When there exists an obviously zero density area between the stacked data and normal data, the DBSCAN algorithm can deal with abnormal data efficiently. However, when stacked data and normal data are closely located, setting the value of two parameters, Minpts and ε, is intractable. It is difficult to set fixed parameters for these intervals whose zero density area between the stacked data and normal data are tiny. Furthermore, DBSCAN will identify all data as noise points in a few intervals with the number of data less than Minpt.

An effective method for stacked data cleaning should be sharper than the *k*-means and DBSCAN. The mean-shift algorithm is a very popular and simple—yet very powerful—unsupervised machine learning algorithm [25]. It is also dependent on densities such as DBSCAN. It tries to find the highest density of data points in the feature space by setting a parameter: bandwidth. In the process of finding this area, every point can be used as a starting point. The procedure of the mean-shift can be summarized as follows.

1. One randomly selects a point in the dataset as the starting center point.
2. One finds out the area with the center and the bandwidth as the radius. It is considered that the points falling in the area belong to the same class, and the number of points in this area is an access frequency of the center point.
3. One takes the starting point as the center point, then calculates the vector from the center to each element in this area and adds all vectors to gain the shift vector.
4. One moves the original center point along the shift vector, and the distance of this movement is the modulus length of the shift vector.
5. Repeat steps 2 to 4 until the variation of the shift vector is small, and then record the center. In this process, all points that appeared in these circles are classified into one category.
6. If the distance between the current class center and other existing class centers is less than the threshold, the two classes will be merged. Otherwise, the current class will be regarded as a new class.
7. Repeat steps 1 to 6 until all points in the dataset are selected.
8. According to the access frequency of each class to each point, the class with the highest access frequency is taken as that of the belonging class of the current point.

In the area formed by the bandwidth, the kernel function is set to the window function, so that the contribution of each point in the area of bandwidth to the center is the same, and the weight of all the points in the bandwidth to the center point is 1. In fact, this contribution is related to the distance from the center to each point. At the same time, the importance of each sample is also different. Based on the above problem, Ref [24] added the sample weight to the window function or changed the window function into a Gaussian kernel function and corrected the shift vector.

In order to enhance the efficiency of the algorithm, one initializes the data by sampling a small number of points in a discrete manner. Compared with the DBSCAN algorithm, the parameter of the mean-shift method is not sensitive to the distribution of the stacked data and normal data. Even if there is no obvious zero density area between the stacked data and normal data, the mean-shift algorithm can effectively distinguish them.

3.4. The Refine Processing of Data Cleaning

Most wind turbines can be presented by a standard power curve model after the above three steps. However, for some wind turbines that work in long-term wind curtailment, they do not keep normal data for the high wind speed intervals. For example, if one takes the No. 6 wind turbine as a case, its power curve is shown in Figure 1d. When the wind speed is more than 17.5 m/s, all point belongs to the stacked data, and the mean-shift algorithm will not be available. In this case, one can firstly filter the abnormal data according to the previous three steps, then find the maximum wind speed $v_{\max}$ corresponding to the normal power curve, and remove the data with wind speeds higher than $v_{\max}$ and power lower than the data corresponding to $v_{\max}$.

4. Case Comparison and Analysis

In order to verify the effectiveness of the proposed algorithm, the QK [9], QD [10], MMO [17], and FA [18] are adopted for comparison. The experimental data of the four wind turbines comes from a prototype wind farm. The power curve of the four wind turbines is shown in Figure 1. Most of the wind power curves analyzed in the literature are similar to that of the No. 5 wind turbines. Consequently, only one type of wind power curve is considered. However, the operating environment of the wind turbine is unstable, thus, the wind power curves will also be similar to those shown in the No. 3, No. 4, and No. 6 wind turbines. The abnormal data of the No. 4 turbine only belong to the scattered data category. The normal data of the No. 6 turbine only accounts for a small part of the total data, and it is very easy to be mistakenly eliminated. The value of k in QK is set to 8. The Minpt of QD is equal to 5 and the ε is 25 kW. The size of all the WPC binary images is set to 432×288, and each data point is set to four pixels, which is consistent with [17,18]. The best structuring element size of the MMO is calculated by the Hu moment. The reference image of WPC can be extracted [17]. For the combined algorithm proposed in this paper, the power interval is divided into 25 kW and the wind speed interval is divided into 0.5 m/s. The contamination of I-Forest is set to 0.1 in every power interval for the No. 3, No. 4, and No. 5 wind turbines, but when the normal data is far less than abnormal data, this proportion needs to be adjusted to 0.05 (e.g., No. 6 wind turbine); then, this parameter should be reduced to a half rate in each wind interval if there exists some scattered data. If the scattered data is removed in the first step or two steps by IF, one uses the mean-shift algorithm to directly clean the remaining stacked data. The bandwidth of the mean-shift can be set to 135. The results of removing the abnormal wind power data corresponding to the above five methods are shown in Figures 4–8. All algorithms are programmed using Python 3.6.5.

From Figures 4 and 5, it can be noted that the QK and QD have a better ability to remove abnormal data for the No. 4 and No. 5 wind turbines. For the No. 3 wind turbine, the quartile method is not good at removing the scattered data closing to the normal curve, and because the data of No. 3 turbine is insufficient in the high wind speed range, the k-means algorithm cannot identify a few of the stacked data in twice clustering. However,

although the QD algorithm can avoid the setting of the k-value and can eliminate some scattered data closing to the curve, it cannot effectively identify the stacked data when the gap between the stacked data and the normal data is not obvious. In addition, the anomalous data in the No. 6 wind turbine is far more than the normal data; the quartile method has eliminated all normal data. Since the normal data are mistakenly excluded, the No. 6 wind turbine no longer uses the *k*-means or DBSCAN-clustering algorithms to eliminate abnormal data again.

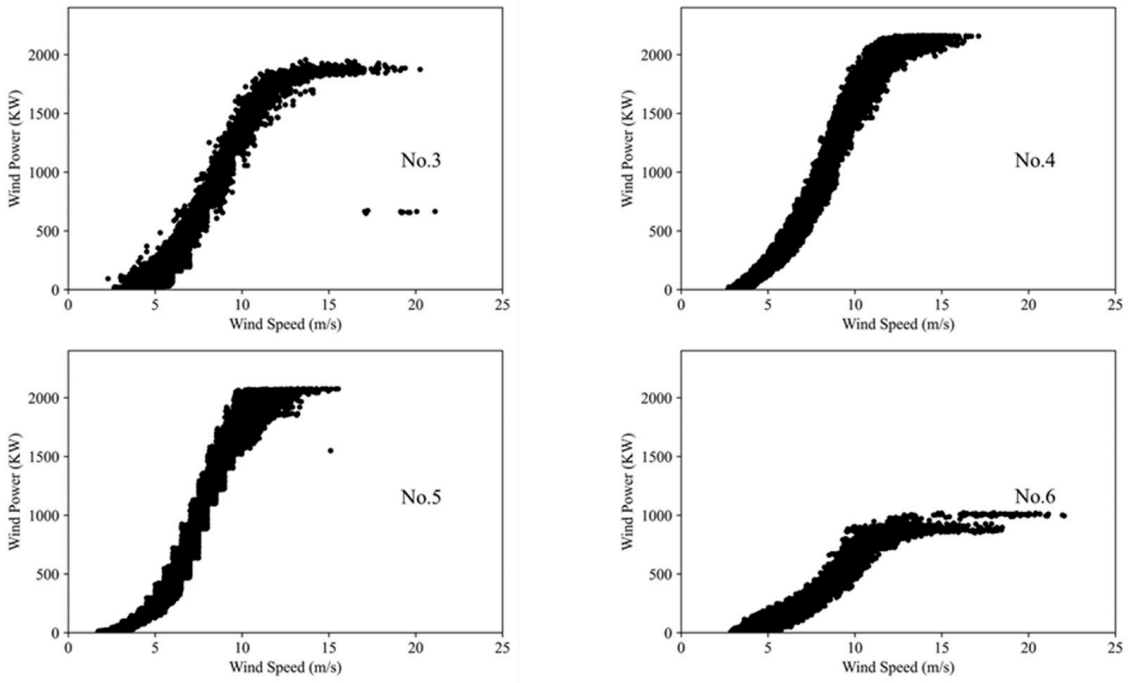

Figure 4. Abnormal data filtering result via QK.

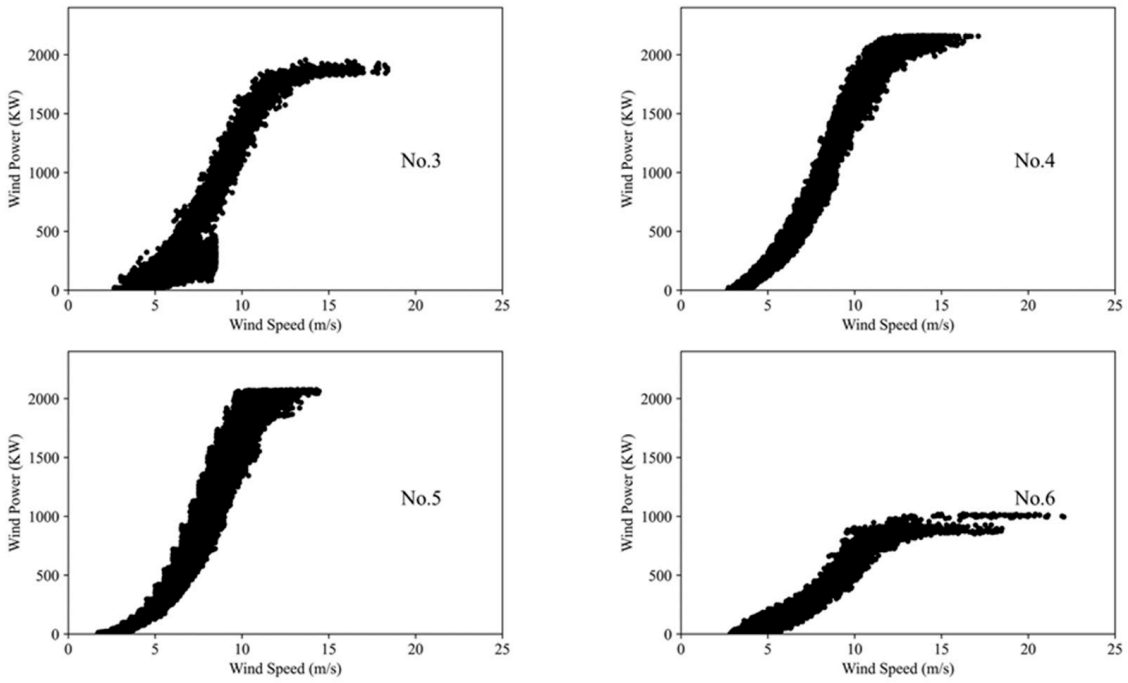

Figure 5. Abnormal data filtering result via QD.

From Figures 6 and 7, it can be found that when the abnormal data with a high density is close to the normal curve, converting wind power curves into images will easily take the abnormal and normal data as a whole. At this time, the result of the MMO method and the FA method will deteriorate. MMO cannot handle abnormal data that is assimilated by normal data into a binary image, and the FA algorithm can also cause severe distortion of the wind power curve. Similarly, the method in Ref [20] cannot clean the wind turbines No. 3, No. 4, and No. 6 as well, since the pixel values of many abnormal data in the feature image will be larger than the normal data. The density of stacked data in these wind turbines is comparable to normal data, thus, the MMO and the FA are invalid in these three cases. Obviously, an image processing method can efficiently remove the scattered data since the size of the scattered data in the image is smaller after that the wind power curve is transformed into a binary image. However, such a method performs poorly in removing stacked data because a higher proportion of stacked data will be tuned into the main part of the image easily and will be retained in the image.

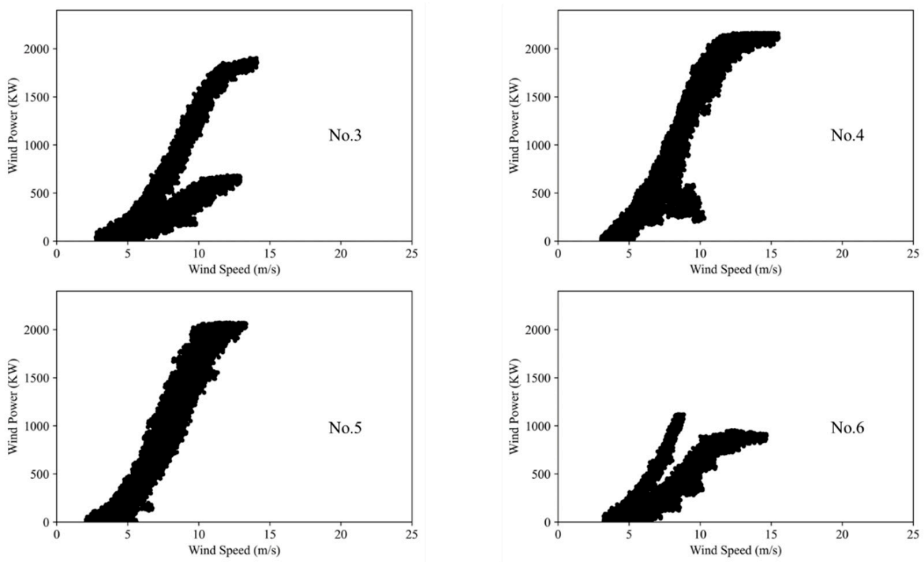

Figure 6. Abnormal data filtering result via MMO.

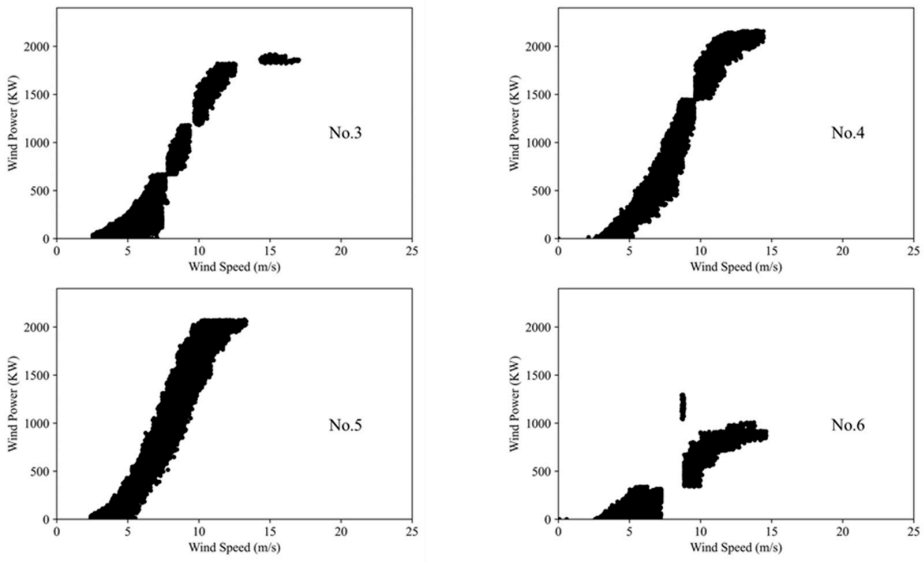

Figure 7. Abnormal data filtering result via FA.

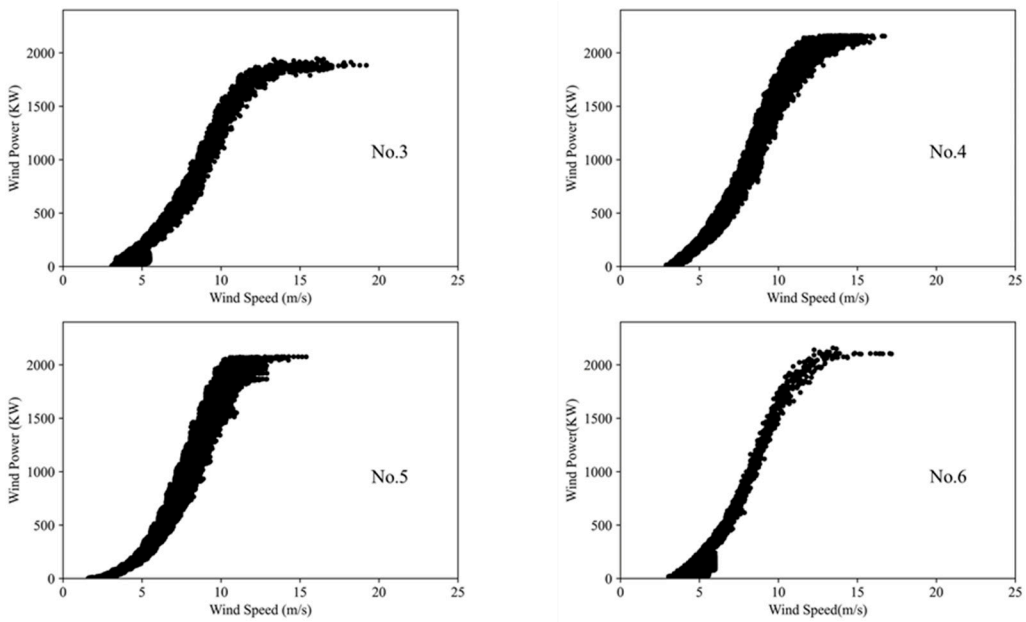

Figure 8. Abnormal data filtering result via proposed method.

Figure 8 is the data cleaning result of the method proposed in this paper. It can be found that the four wind turbines, especially the No. 3 and No. 6 turbines, have achieved relatively ideal results. Comparing Figures 4, 5 and 8, it is clear that a clustering algorithm can effectively eliminate the stacked data as compared to an image processing one. However, it is advisable to eliminate the scattered data on the wind power curve before a clustering algorithm positively confirms the necessity of introducing the Isolation Forest, as reported in this paper. Moreover, the parameters of the clustering algorithm play a vital role in the algorithm's performance. Additionally, compared with the processing results of the No. 3 wind turbine via QK, QD, and the proposed method, the mean-shift is more sensitive than the *k*-means and DBSCAN to the classification of stacked data and normal data.

The efficiency and results of the five algorithms are compared in Table 1. The performance parameters include the recorded data amount of each wind turbine, the deletion rate of the outliers R (%) and the calculation time T (s). Metric R (%) is defined as the deletion rate of the removed data on the raw data. Generally, a high R (%) means that more abnormal data have been successfully filtered. Compared with the other methods, it is easy to observe that the proposed method filters more abnormal data, irrespective of either the stacked or scattered abnormal ones.

Table 1. Experimental results of five algorithms.

Wind Turbine	Data Amount	Proposed		QK		QD		MMO		FA	
		R (%)	T (s)	R (%)	T (s)	R (%)	T (s)	R (%)	T (s)	R (%)	T (s)
No. 3	38,995	58.9	44.7	49.8	0.83	38.5	0.86	32.1	2.86	45.1	1.34
No. 4	44,335	38.6	45.6	33.4	0.24	33.4	0.24	30.6	2.91	33.0	1.29
No. 5	50,962	47.1	47.2	47.6	0.71	41.6	0.67	40.5	2.93	39.6	1.44
No. 6	45,592	64.6	38.1	44.1	0.23	44.1	0.23	39.2	2.88	50.3	1.47

It should be noted that an extremely high R (%) may result in an over-cleaning problem; and in such cases, the approach introduced in [26] can be applied.

To sum up, the proposed algorithm can not only handle conventional abnormal wind power data, but it can also provide a reliable power curve, even in serious wind curtailment. Image algorithms are not suitable for processing high-density stacked abnormal data, as their ability to identify outliers will become weak. If the scattered data are distributed near

the normal power curve, QK and QD may not recognize them. In addition, unreasonable parameter settings may lead to a misidentification phenomenon, which will decrease the recognition accuracy.

5. Conclusions

In order to obtain accurate wind power forecasting results, this paper introduces an efficient algorithm by combining the Isolation Forest and mean-shift to remove abnormal data. When using the Isolation Forest to process abnormal data in a specific interval, the abnormal proportion can be set according to the specific situation of the wind power curve, facilitating it a certain flexibility. Moreover, the mean-shift algorithm can effectively eliminate the remaining stacked data. Consequently, the proposed algorithm integrates the advantages of both the Isolation Forest and a mean-shift approach. The numerical results of filtering the abnormal data of four wind turbines have demonstrated that the proposed algorithm can not only handle conventional abnormal wind power data, but can also provide a reliable power curve, even in serious wind curtailment. It should be noted that some over-cleaning problems, as commonly encountered by the existing filtering algorithms, including the proposed one, may result in a degradation in the algorithm performance. The authors will strive to solve this issue in their further work in this direction.

Author Contributions: Methodology, W.W.; Supervision, S.Y.; Visualization, Y.Y.; Writing—original draft, W.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Wan, C.; Xu, Z.; Pinson, P.; Dong, Z.; Wong, K. Probabilistic forecasting of wind power generation using extreme learning machine. *IEEE Trans. Power Syst.* **2014**, *29*, 1033–1044. [CrossRef]
2. Wan, C.; Xu, Z.; Pinson, P.; Dong, Z.; Wong, K. Optimal prediction intervals of wind power generation. *IEEE Trans. Power Syst.* **2014**, *29*, 1166–1174. [CrossRef]
3. Long, H.; Zhang, C.; Geng, R.; Wu, Z.; Gu, W. A combination interval prediction model based on biased convex cost function and auto-encoder in solar power prediction. *IEEE Trans. Sustain. Energy* **2021**, *12*, 1561–1570. [CrossRef]
4. Zhao, C.; Wan, C.; Song, Y. Operating reserve quantification using prediction intervals of wind power: An integrated probabilistic forecasting and decision methodology. *IEEE Trans. Power Syst.* **2021**, *36*, 3701–3714. [CrossRef]
5. Zhao, C.; Wan, C.; Song, Y. An adaptive bilevel programming model for nonparametric prediction intervals of wind power generation. *IEEE Trans. Power Syst.* **2020**, *35*, 424–439. [CrossRef]
6. Zhao, Y.; Ye, L.; Pinson, P.; Tang, Y. Correlation-constrained and sparsity-controlled vector autoregressive model for spatio-temporal wind power forecasting. *IEEE Trans. Power Syst.* **2018**, *33*, 5029–5040. [CrossRef]
7. Shen, X.; Fu, X.; Zhou, C. A combined algorithm for cleaning abnormal data of wind turbine power curve based on change point grouping algorithm and quartile algorithm. *IEEE Trans. Sustain. Energy* **2019**, *10*, 46–54. [CrossRef]
8. Xiang, L.; Yang, X.; Hu, A.; Su, H.; Wang, P. Condition monitoring and anomaly detection of wind turbine based on cascaded and bidirectional deep learning networks. *Appl. Energy* **2022**, *305*, 117925. [CrossRef]
9. Zhao, Y.; Ye, L.; Zhu, Q. Characteristics and processing method of abnormal data clusters caused by wind curtailments in wind farms. *Autom. Electr. Power Syst.* **2014**, *38*, 39–46.
10. Zhao, Y.; Hu, Q.; Srinivasan, D.; Wang, Z. Data-driven correction approach to refine power curve of wind farm under wind curtailment. *IEEE Trans. Sustain. Energy* **2018**, *9*, 95–105. [CrossRef]
11. Zheng, L.; Hu, W.; Min, Y. Raw wind data preprocessing: A data-mining approach. *IEEE Trans. Sustain. Energy* **2015**, *6*, 11–19. [CrossRef]
12. Zhou, Q.; Ma, Y.; Lv, Q. Abnormal data processing of wind turbine based on combined algorithm and class center imputation. In Proceedings of the 2021 International Conference on Power System Technology (POWERCON), Haikou, China, 8–9 December 2021.
13. Khazaei, S.; Ehsan, M.; Soleymani, S.; Mohammadnezhad-Shourkaei, H. A high-accuracy hybrid method for short-term wind power forecasting. *Energy* **2022**, *238*, 122020. [CrossRef]

14. Hu, Y.; Qiao, Y.; Liu, J.; Zhu, H. Adaptive confidence boundary modeling of wind turbine power curve using SACADA data and its application. *IEEE Trans. Sustain. Energy* **2018**, *10*, 1330–1341. [CrossRef]
15. Xi, Y.; Lu, Z.; Ying, Q.; Yong, M.; O'Malley, M. Identification and correction of outliers in wind farm time series power data. *IEEE Trans. Power Syst.* **2016**, *31*, 4197–4205.
16. Guo, P.; Infield, D. Wind turbine power curve modeling and monitoring with Gaussian process and SPRT. *IEEE Trans. Sustain. Energy* **2020**, *11*, 107–115. [CrossRef]
17. Long, H.; Sang, L.; Wu, Z.; Gu, W. Image-based abnormal data detection and cleaning algorithm via wind power curve. *IEEE Trans. Sustain. Energy* **2020**, *11*, 938–946. [CrossRef]
18. Wang, Z.; Wang, L.; Huang, C. A fast abnormal data cleaning algorithm for performance evaluation of wind turbine. *IEEE Trans. Instrum. Meas.* **2021**, *70*, 5006512. [CrossRef]
19. Su, Y.; Chen, F.; Liang, G.; Wu, X.; Gan, Y. Wind power curve data cleaning algorithm via image thresholding. *Proc. Int. Conf. Robot. Biomim.* **2019**, 1198–1203.
20. Liang, G.; Su, Y.; Chen, F.; Long, H.; Song, Z. Wind power curve data cleaning by image thresholding based on class uncertainty and shape dissimilarity. *IEEE Trans. Sustain. Energy* **2021**, *12*, 1383–1393. [CrossRef]
21. Available online: https://github.com/AmangAris/Abnormal-data-identification-and-cleaning-of-wind-turbine (accessed on 6 June 2022).
22. Liu, F.T.; Ting, K.M.; Zhou, Z.H. Isolation forest. In Proceedings of the 2008 Eighth IEEE International Conference on Data Mining, Pisa, Italy, 15–19 December 2008; pp. 413–422.
23. Li, C.; Guo, L.; Gao, H.; Li, Y. Similarity-Measured isolation forest: Anomaly detection method for machine monitoring data. *IEEE Trans. Instrum. Meas.* **2021**, *70*, 3512512. [CrossRef]
24. Fen, C.; Zhu, S.; Zhu, Z.; Sun, M. Comparative study on detection methods of abnormal wind power data. *Adv. Technol. Electr. Eng. Energy* **2021**, *40*, 55–61.
25. Cheng, Y. Mean Shift, Mode Seeking, and Clustering. *IEEE Trans. Pattern Anal. Mach. Intell.* **1995**, *17*, 790–799. [CrossRef]
26. Yesilbudak, M. Implementation of Novel Hybrid Approaches for Power Curve Modeling of Wind Turbines. *Energy Convers. Manag.* **2018**, *171*, 156–169. [CrossRef]