energies

Article

Abnormal Data Cleaning Method for Wind Turbines Based on Constrained Curve Fitting

Xiangqing Yin, Yi Liu, Li Yang and Wenchao Gao *

School of Mechanical Electronic and Information Engineering, China University of Mining and Technology-Beijing, Beijing 100083, China
* Correspondence: gaowc@cumtb.edu.cn

Abstract: With the increase of the scale of wind turbines, the problem of data quality of wind turbines has become increasingly prominent, which seriously affects the follow-up research. A large number of abnormal data exist in the historical data recorded by the wind turbine Supervisory Control And Data Acquisition (SCADA) system. In order to improve data quality, it is necessary to clean a large number of abnormal data in the original data. Aiming at the problem that the cleaning effect is not good in the presence of a large number of abnormal data, a method for cleaning abnormal data of wind turbines based on constrained curve fitting is proposed. According to the wind speed-power characteristics of wind turbines, the constrained wind speed-power curve is fit with the least square method, and the constrained optimization problem is transformed into an unconstrained optimization problem by using the external penalty function method. Data cleaning was performed on the fitted curve using an improved 3-σ standard deviation. Experiments show that, compared with the existing methods, this method can still perform data cleaning well when the historical wind turbine data contains many abnormal data, and the method is insensitive to parameters, simple in the calculation, and easy to automate.

Keywords: data cleaning; wind turbine; wind power; constrained curve fitting

Citation: Yin, X.; Liu, Y.; Yang, L.; Gao, W. Abnormal Data Cleaning Method for Wind Turbines Based on Constrained Curve Fitting. *Energies* **2022**, *15*, 6373. https://doi.org/10.3390/en15176373

Academic Editor: Davide Astolfi

Received: 10 July 2022
Accepted: 30 August 2022
Published: 31 August 2022

1. Introduction

Currently, with the development of the world economy and society, the emphasis on environmental protection is getting higher and higher, and the new energy power generation industry has also developed rapidly. As stated in the "Global Green Energy Status Report", 26.2% of the energy in electricity production is green energy, and wind energy accounts for 5.5% of the green energy [1], so the collection and use of wind energy has become particularly important. The number of studies on wind power generation is also increasing. The historical operation data and operation data of wind turbine are regarded as the basis of wind power research. Through the historical operation data of wind turbines, not only is output prediction possible [2,3], but also condition monitoring [4,5] and fault diagnosis [6–8]. The Supervisory Control And Data Acquisition system (SCADA) of wind turbines records a large amount of wind historical operation data and process control information, including wind speed, power, wind direction, rotation speed, etc. It is of great significance to study the operation law of wind turbines and predict early failures. However, due to human errors, sensor failures, or communication failures, SCADA data may contain a large amount of abnormal data, which reduces data quality and hinders subsequent research. Therefore, the data quality of wind turbines has also been paid more and more attention [9,10]. Data cleaning is of great significance for improving data quality and promoting follow-up research.

In order to improve the quality of data, many methods have been proposed to clean abnormal data. According to the method used for data cleaning, it can be divided into the following categories.

(1) Image method. The basic idea of the image method is to convert scattered data into digital images and transform the data cleaning problem into an image segmentation problem. Huan Long et al. [11] proposed an abnormal data cleaning algorithm based on 3D images. Su Y et al. [12] and Liang G et al. [13] used an image thresholding algorithm to identify anomalies. Wang Z et al. [14] propose an efficient acceleration algorithm that can convert data into images for cleaning. The disadvantage of the image method is that the required computing resources are too large.

(2) Power curve modeling method. The power curve modeling method is to establish a wind speed-power curve model through a series of methods, compare the real data with the power curve model, and then clean out abnormal data. The methods include quantile power curve [15], interval extreme probability density [16], maximum likelihood estimation [17], Artificial Neural Network (ANN) algorithm [18], etc. Based on the ideal curve, Joon-Young Park et al. [19] used a monitoring power curve to automatically calculate the limit of the power curve value method. Yongning Zhao et al. [20] proposed an algorithm combining the quartile algorithm and the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) to optimize the power curve, in which the quartile algorithm was used to eliminate sparse abnormal data, and DBSCAN was used to eliminate accumulated abnormal data. According to different wind characteristics, Yang Mao et al. [21] used the Copula function to obtain the probability power curve and combined the time series characteristics of the abnormal data to summarize three types of abnormal data, and established the abnormal data identification model, which improved the modeling accuracy. The disadvantage of the power curve modeling method is that when there are a large number of abnormal data, the abnormal data will greatly affect the accuracy of the established power curve.

(3) Statistical methods. Statistical methods compare the statistical value of data or data in each interval with a preset threshold to achieve the purpose of data cleaning. Statistical methods include sample entropy [22], cloud segmentation optimal entropy [23], bin algorithm [24,25], quartile method [26], etc. Lou Jianlou et al. [27] used the optimal intra-group variance method for data cleaning, which is good at dealing with abnormal data with low power in the wind speed range. Wang S et al. [28] adopted the combination of 3σ-median criterion to effectively identify abnormal data points in the data. Tao L et al. [29] use the gray relational algorithm and the support vector regression algorithm to effectively solve the problem of dimensional explosion. There are some algorithms that divide abnormal data into multiple types and use different algorithms according to the characteristics of the types [30,31]. The statistical data generated by statistical methods will be affected by abnormal data, and there is a problem of difficulty in threshold selection.

In summary, in order to improve the data quality, it is necessary to effectively clean a large number of abnormal data in the original data. Therefore, the abnormal data cleaning method for wind turbines based on constrained curve fitting is proposed. Aiming at the problems existing in the existing methods, including a large amount of calculation, difficult parameter selection, and poor cleaning effect when a large number of abnormal data exist, the following work is done in this paper. (1) Firstly, the wind speed-power scattergram of historical data of wind turbines is analyzed, and each wind speed interval and power interval is preprocessed by quartile. (2) According to the wind speed-power characteristics and the ideal curve, the constraint function is set, and the least square method used to fit the wind speed-power data constrained curve. The constrained optimization problem is transformed into an unconstrained optimization problem by using the exterior penalty function method. (3) The modified 3-σ standard deviation is used for the fitted curve to extract normal data. In the remaining data, the data within the rated wind speed range are classified as normal data, and the rest are classified as abnormal data. (4) Taking the real data of the wind turbine as a sample, the experiments show that the method can still perform data cleaning well in the presence of a large number of abnormal data, the method is not sensitive to parameters, and the method is easy to calculate and realize automation.

This article has five sections. Section 2 discusses the wind speed-power characteristics of wind turbines, shows the ideal wind speed-power curve, and defines what is abnormal data. In Section 3, the methods used are briefly introduced, including the quartile algorithm, the constrained least squares curve fitting algorithm, the external penalty function method, and the 3-σ algorithm. An abnormal data cleaning process for wind turbines based on constrained curve fitting is established. In Section 4, the effectiveness of the algorithm is verified in experiments and the robustness of the algorithm is discussed. Entropy and hyper entropy are used as indicators to evaluate the stability of the cleaned data. Compared with the traditional algorithm, it shows that the method is insensitive to parameters and has the advantages of small calculation amount. Section 5 is the conclusion of this paper.

2. Wind Speed-Power Characteristics

Wind turbines have the function of converting wind energy into electric energy. The principle of wind power generation is that wind turbine blades rotate through wind, and the wind energy is converted into mechanical work to drive the rotor to rotate, so that the generator generates electricity. The wind speed-power curve of the wind turbine is a function representing the performance of the wind turbine and represents the relationship between the output power P_0 of the wind turbine and the wind speed v. The ideal wind speed-power curve can be written in this form [32]:

$$P_0(v) = \frac{1}{2}\rho A v^3 C_{P_0}(v) \tag{1}$$

where ρ is the air density; A is the area swept when the fan blade rotates; and $C_{P_0}(v)$ is the ideal power index, which can be regarded as a function of wind speed.

In the wind turbine, for the purpose of protecting the wind turbine to generate electricity continuously and stably, and avoid accidents such as tower collapse when the wind speed is too high, the function of the actual output power P and the wind speed v can be seen as:

$$P = \begin{cases} 0 & v < v_{\text{in}}, v > v_{\text{out}} \\ P_0(v) & v_{\text{in}} \leqslant v < v_{\text{r}} \\ P_{\text{r}} & v_{\text{r}} \leqslant v < v_{\text{out}} \end{cases} \tag{2}$$

where v_{in} is the cut-in wind speed; v_{out} is the cut-out wind speed; v_{r} is the rated wind speed; and P_{r} is the rated power.

The cut-in wind speed of the wind turbine is the lowest wind speed at which the external power transmission starts, and the cut-out wind speed is the highest wind speed for grid-connected power generation. When the wind speed is not within the range of v_{in} and v_{out}, it is not integrated into the grid, and the output power is 0. When the wind speed is within the range of cut-in wind speed and rated wind speed, the wind turbine generates power stably, and when the wind speed changes, the output power changes accordingly. When the wind speed v reaches the rated wind speed v_{r}, the wind turbine reaches the preset maximum output power, that is, the rated power. When the wind speed is within the range of rated wind speed and cut-out wind speed, wind turbine keeps the rated power unchanged.

However, due to the influence of unstable wind speed, turbulence, and yaw error, there is a large difference between the ideal output power and the actual output power, and it is meaningless to directly apply the ideal wind speed-power curve. However, the ideal wind speed-power curve reveals the relationship between wind speed and output power, and the purpose of identifying abnormal data can be reached by fitting the wind power curve. The actual wind speed-power scatter and ideal wind speed-power curve of a wind turbine located in eastern China are shown in Figure 1.

In the wind turbine operation data, the data that do not conform to the wind speed-power characteristics are called abnormal data. It can be seen from Figure 1 that there are mainly two types of abnormal data within the range of cut-in wind speed and cut-out wind speed: (1) upper abnormal data: data with low wind speed and high power; (2) limited

power abnormal data: high wind speed with low output power. In actual operation, a large number of abnormal data may be generated and the distribution of different abnormal data may affect the effect of data cleaning.

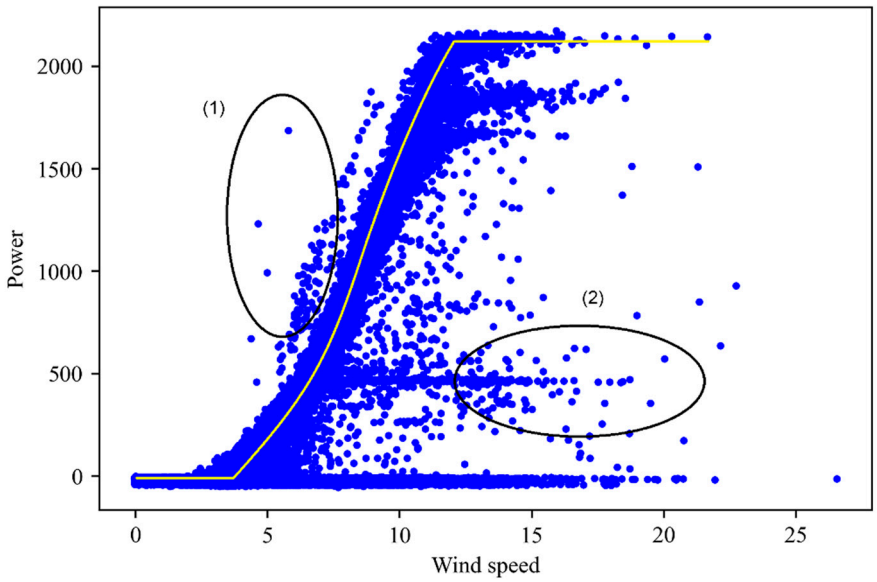

Figure 1. Actual wind speed-power scatter point and ideal wind speed-power curve.

3. Abnormal Data Cleaning Method for Wind Turbines Based on Constrained Curve Fitting

The monitoring and data acquisition system (SCADA) of the wind turbine records a large amount of historical wind operation data. According to Formula (2), the data part of the wind speed greater than the cut-in wind speed and less than the cut-out wind speed is selected for data cleaning. The process of the abnormal data cleaning method for wind turbines based on constrained curve fitting is shown in Figure 2.

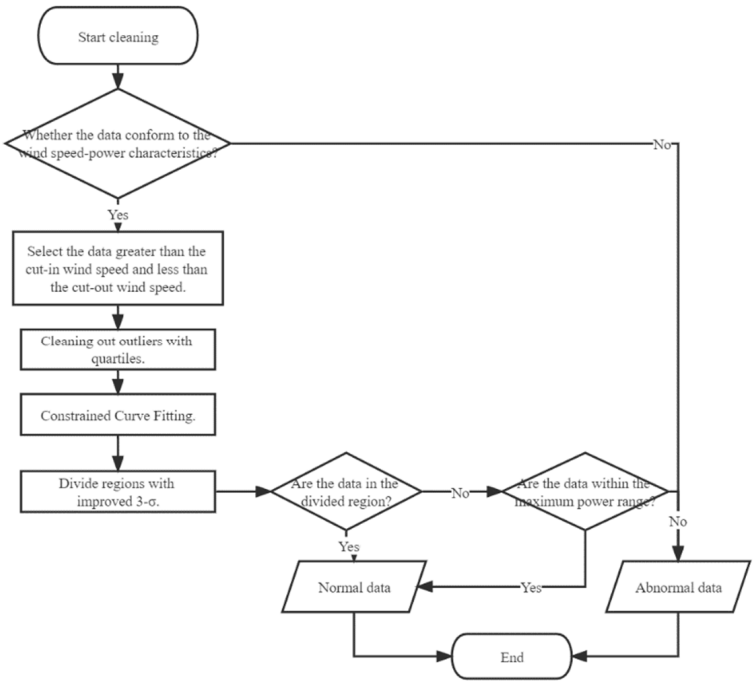

Figure 2. The process of the abnormal data cleaning method for wind turbines based on constrained curve fitting.

3.1. Quartile Outlier Data Detection Algorithm

The quartile is a type of statistical quantile. All data are arranged from small to large. The number in the quarter is called the lower quartile Q_1, and the number in the third quarter is called the upper quartile Q_3. The upper quartile minus the lower quartile is called interquartile range $IQR = Q_3 - Q_1$. For any data, if it is less than $Q_1 - 1.5IQR$, it can be considered that the data value is too small; if it is greater than $Q_3 + 1.5IQR$, it can be considered that the data value is too large. Data values that are too small or too large can be called outliers, and the quartile algorithm can be used to identify and clean outliers.

3.2. Constrained Least Square Curve Fitting Algorithm

Observing the ideal wind speed-power curve in Figure 1, it is found that the ideal wind speed-power curve of the wind turbine is similar to the sigmoid function $S(x) = \frac{1}{1+e^{-x}}$. The graph of the sigmoid function and its derivative is shown in Figure 3.

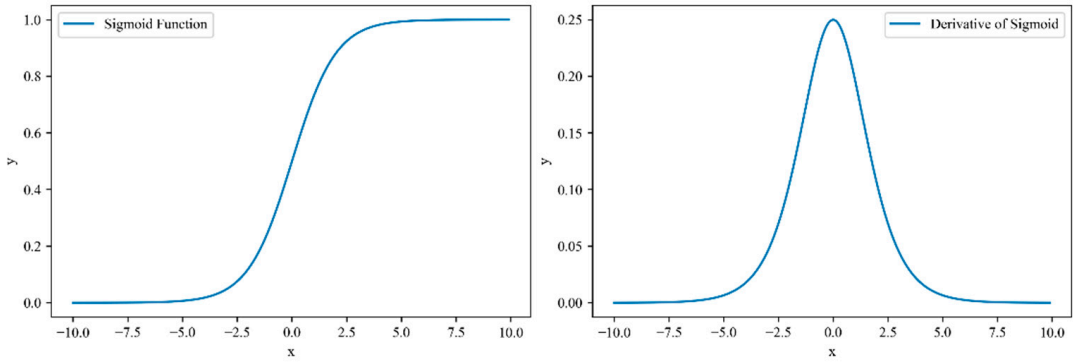

Figure 3. Sigmoid function and the derivative of the Sigmoid function image.

Therefore, the Sigmoid-like function can be used to fit the data processed by the quartile method, and the fitting function can be expressed as:

$$h\left(\vec{x}, v\right) = \frac{x_0}{x_1 + e^{-(x_2 v + x_3)}} \tag{3}$$

$$\frac{\partial h\left(\vec{x}, v\right)}{\partial v} = \frac{x_0 x_2 e^{-(x_2 v + x_3)}}{\left(x_1 + e^{-(x_2 v + x_3)}\right)^2} \tag{4}$$

Among them, v is the wind speed; $\vec{x}(x_0, x_1, x_2, x_3)$ is the parameter; $h\left(\vec{x}, v\right)$ is the output power of the wind speed v under the condition of the parameter $\vec{x}(x_0, x_1, x_2, x_3)$; and $\frac{\partial h\left(\vec{x}, v\right)}{\partial v}$ is the partial derivative of the wind speed v with respect to $h\left(\vec{x}, v\right)$.

For any real data (v_i, P_i), the value of $h\left(\vec{x}, v_i\right) - P_i$ is called the residual, and the purpose of the least square method is to find the parameter $\vec{x}(x_0, x_1, x_2, x_3)$ that minimizes the sum squared residual of all the data in the sample. Denoting the sum squared residual as the objective function $f\left(\vec{x}\right)$, the objective of unconstrained curve fitting can be expressed as:

$$minf\left(\vec{x}\right) = min\sum_{i}^{n}\left(h\left(\vec{x}, v_i\right) - P_i\right)^2 = min\sum_{i}^{n}\left(\frac{x_0}{x_1 + e^{-(x_2 v_i + x_3)}} - P_i\right)^2 \tag{5}$$

Since the dataset contains a large number of abnormal data, if the Formula (5) is directly used for fitting, the curve fitting result will be shifted due to the influence of the abnormal data, so the curve fitting process needs to be constrained. Figure 4 shows the unconstrained curve fitting error under the influence of abnormal data.

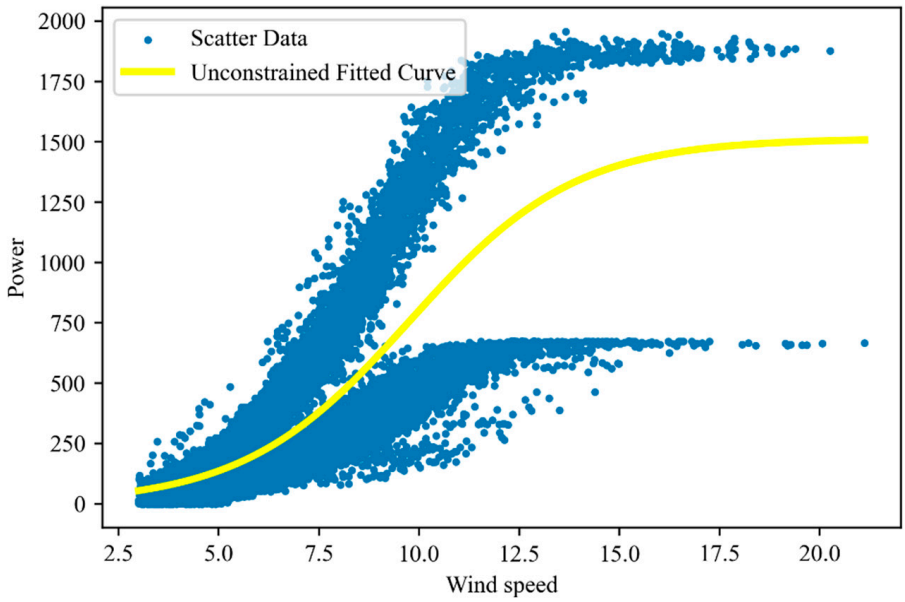

Figure 4. The unconstrained curve fitting error under the influence of abnormal data.

Due to communication errors, sensor errors, etc., the actual maximum power $P_{\max}$ of the wind turbine may not be exactly equal to the rated power P_r, so the error value is set as $\Delta = 5\% \times P_r$; it can be considered that the actual maximum power is within the error range of the rated power, that is, $P_r - \Delta \le P_{\max} \le P_r + \Delta$. Let any data in the dataset be (v_i, P_i), the set of this part of the data can be expressed as $S = \{ (v_i, P_i) \mid P_r - \Delta \le P_i \le P_r + \Delta,\ v_{in} \le v_i \le v_{out} \}$. In the case of not knowing the rated wind speed, it can be considered that the power corresponding to the average value of the wind speed of the data whose power is within the error range of the rated power is the maximum power. The partial derivative corresponding to the mean wind speed should be a small value. For Formula (3), x_0, x_1, x_2 should all be positive numbers. Therefore, constrained curve fitting can be transformed into a constrained optimization problem:

$$\begin{aligned} min \quad & f\left(\vec{x}\right) \\ s.t. \quad & g_1\left(\vec{x}\right) = h\left(\vec{x}, v_{mean}\right) - P_r + \Delta - \delta \ge 0 \\ & g_2\left(\vec{x}\right) = P_r + \Delta - h\left(\vec{x}, v_{mean}\right) - \delta \ge 0 \\ & g_3\left(\vec{x}\right) = \alpha - \frac{\partial h\left(\vec{x}, v\right)}{\partial v}|_{v_{mean}} - \delta \ge 0 \\ & g_4\left(\vec{x}\right) = x_0 - \delta \ge 0 \\ & g_5\left(\vec{x}\right) = x_1 - \delta \ge 0 \\ & g_4\left(\vec{x}\right) = x_2 - \delta \ge 0 \end{aligned} \tag{6}$$

Among them, v_{mean} is the average value of the wind speed within the error range of the rated power, which can be expressed as $v_{mean} = \left\{ \frac{\sum_i^n v}{n} | (v_i, P_i) \in S \right\}$; α is the upper limit of the partial derivative when the wind speed is v_{mean}; and δ is the constraint margin, which ensures that the optimal solution is close to the feasible point.

3.3. Exterior Penalty Function Method for Solving the Constrained Optimization Problem

The exterior penalty function method transforms constrained optimization problem into unconstrained optimization problem by setting a penalty function. The penalty function can be expressed as:

$$\varphi\left(\vec{x}, M^{(k)}\right) = f\left(\vec{x}\right) + M^{(k)} \sum_i^6 min\left(g_i\left(\vec{x}\right), 0\right)^2 \tag{7}$$

Among them, $M^{(k)}$ represents the exterior penalty factor at the k-th iteration, which is an increasing sequence of positive numbers. The larger the $M^{(k)}$, the more severe the penalty.

Therefore, Formula (6) can be transformed into an unconstrained optimization problem:

$$min\ \varphi\left(\vec{x}, M^{(k)}\right) \tag{8}$$

When all constraints $g\left(\vec{x}\right)$ satisfy the constraints, $min\left(g\left(\vec{x}\right), 0\right)^2 = 0$, then the penalty function is equal to the objective function, and solving the penalty function is equivalent to solving the objective function. When any constraint $g_i\left(\vec{x}\right)$ does not satisfy the constraint, $min\left(g_i\left(\vec{x}\right), 0\right)^2 = {g_i}^2$, the penalty function is penalized, and the more $g_i\left(\vec{x}\right)$ does not satisfy the constraint, the more severe the penalty. The algorithm flow of the exterior penalty function method is shown in Algorithm 1.

Algorithm 1 Exterior Penalty Function Method

Input	$\vec{x}$: initial parameters; $M^{(1)}$: initial exterior penalty factor; c: amplification factor; ε_1, ε_2: precision; R: penalty factor control factor
Output	$\vec{x}^*$: optimal parameters
1:	$k = 1$.
2:	Solve the unconstrained optimization problem $min\ \varphi\left(\vec{x}, M^{(k)}\right)$, and get $\vec{x}^*$.
3:	If $min\left\{g_i\left(\vec{x}^*\right) \middle\| i = 1, 2 \cdots m\right\} \geq \varepsilon_1$, go to step 7, otherwise go to step 4.
4:	If $M^{(k)} > R$ *and* $\left\|\vec{x}^* - \vec{x}\right\| \leq \varepsilon_2$, go to step 7, otherwise go to step 5.
5:	$\vec{x} = \vec{x}^*$, $M^{(k+1)} = cM^{(k)}$.
6:	$k = k + 1$, go to step 2.
7:	output $\vec{x}^*$.

The exterior penalty function is defined outside the feasible region and gradually approaches the optimal solution, so there is no requirement for the value of the initial parameter $\vec{x}$. The exterior penalty factor $M^{(k)}$ is gradually increased by the amplification factor c, and generally c is 5–10. When the penalty factor $M^{(k)}$ is too small, the penalty effect is weak and the number of iterations is too large; when the penalty factor $M^{(k)}$ is too large, the penalty effect is strong, the numerical solution is difficult, and the optimization may fail [33]. Therefore, setting penalty factor control factor R. ε_1 and ε_2 are generally selected from 10^{-3}–10^{-4}. If $min\left\{g_i\left(\vec{x}^*\right) \middle\| i = 1, 2 \cdots m\right\} \geq \varepsilon_1$, then $\vec{x}^*$ is close to the constraint boundary, stop iterate.

3.4. Improved 3-σ Data Cleaning Method

The optimal parameter $\vec{x}^*$ is obtained by solving the exterior penalty function and brought into the fitting function $h\left(\vec{x}^*, v\right)$ to obtain the fitting curve. For any wind speed interval U_v^i, find the standard deviation of its power. In the traditional 3-σ method, whether the data are within 3 times the standard deviation of the power is directly used as the criterion to distinguish normal data from abnormal data. In Formula (6), the constraint condition $g_1\left(\vec{x}\right)$ can be regarded as the distance from the theoretical maximum power to the lower limit of the rated power error range, and $g_2\left(\vec{x}\right)$ can be regarded as the distance from the theoretical maximum power to the upper limit of the rated power error range. Therefore, the power upper limit ${P_{high}}_i^j$ and the power lower limit ${P_{low}}_i^j$ of any data point $\left(v_i^j, P_i^j\right)$ in any wind speed interval U_v^i can be expressed as:

$$\begin{aligned} P_{high_i}^{\ j} &= h\left(\vec{x}^*, v_i^j\right) + \frac{3\sigma^i g_2\left(\vec{x}^*\right)}{g_1\left(\vec{x}^*\right)+g_2\left(\vec{x}^*\right)} \\ P_{low_i}^{\ j} &= h\left(\vec{x}^*, v_i^j\right) - \frac{3\sigma^i g_1\left(\vec{x}^*\right)}{g_1\left(\vec{x}^*\right)+g_2\left(\vec{x}^*\right)} \end{aligned} \tag{9}$$

Among them, $\vec{x}^*$ is the optimal parameter obtained by the exterior penalty function method; σ^i is the standard deviation of the power in the wind speed interval U_v^i.

For any data point $\left(v_i^j, P_i^j\right)$ in any wind speed interval U_v^i, when P_i^j is not within the range of $\left[P_{low_i}^{\ j}, P_{high_i}^{\ j}\right]$, it can be considered as abnormal data. The process of the abnormal data cleaning method for wind turbines based on constrained curve fitting is shown in Algorithm 2.

Algorithm 2 Abnormal Data Cleaning Method for Wind Turbines Based on Constrained Curve Fitting

Input:	original dataset: $D = \{(v_1, P_1), (v_2, P_2), \cdots (v_n, P_n)\}$
Output:	normal dataset: D_n
1:	$D_n = \varnothing$
2:	Get the cut-in wind speed v_{in}, cut-out wind speed v_{out}, and rated power P_r from dataset D
3:	$D_0 = \{(v, P) \in D \vert P = 0 \ \wedge (v \le v_{in} \vee v \ge v_{out})\}$ $D_1 = \{(v, P) \in D \vert v_{in} < v < v_{out}\}$ //Select runtime data
4:	$D_n = D_n \cup D_0$
5:	$D_2 = quartile(D_1)$ //Quartile method to remove abnormal data
6:	$\vec{x}^* = sumt(D_2)$ //Exterior penalty function method solves constrained fitting
7:	$\forall (v, P) \in D_2$, calculate its power upper limit P_{high} and power lower limit P_{low}.
8:	$D_3 = \left\{(v, P) \in D_2 \middle\vert P_{low} \le P \le P_{high}\right\}$ //Improved 3-σ method data cleaning
9:	$D_n = D_n \cup D_3$
10:	Bring the maximum wind speed v_{max} of the dataset D_3 into the fitting function to get the actual power maximum value $P_{max} = h\left(\vec{x}^*, v_{max}\right)$.
11:	$D_4 = \left\{(v, P) \in D_1 \middle\vert v_{max} < v < v_{out} \wedge 0.95\, P_{max} \le P \le 1.05 P_{high}\right.$ //Handling the wind speed cut-off part
12:	$D_n = D_n \cup D_4$
13:	output D_n

4. Experimental Validation and Analysis

Using real data, establish a wind turbine abnormal data cleaning model to realize automatic cleaning of abnormal data.

4.1. Dataset Description

The dataset comes from a wind farm in eastern China. It records the SCADA operation data of 12 wind turbines for one year, which is recorded every 10 min. The information contained in the data is shown in Table 1. Different wind turbines face different problems. Three wind turbines with typical abnormal data distribution are selected as #1, #2, and #3. The original data wind speed-power scatter diagram of the three wind turbines is shown in Figure 5.

In Figure 5, the #1 wind turbine contains some upper abnormal data and a small amount of limited power abnormal data, and the abnormal data account for a small proportion; the #2 wind turbine contains lots of limited power abnormal data, and a large amount of abnormal data changes the statistical characteristics of the original data, the proportion of abnormal data and normal data is equal; #3 wind turbine also contains a large number of limited power abnormal data, but abnormal data account for a larger proportion than normal data. The abnormal data cleaning model for wind turbines needs to deal with these three different abnormal data distributions at the same time.

Table 1. SCADA data information.

Name	Unit
Wind Number	
Time Stump	
Wind Speed	m/s
Power	kW
Rotor Speed	r/min
Wheel Diameter	m
Wind Cut-in	m/s
Wind Cut-out	m/s
Rated Power	kW
Rotor Speed Range	r/min

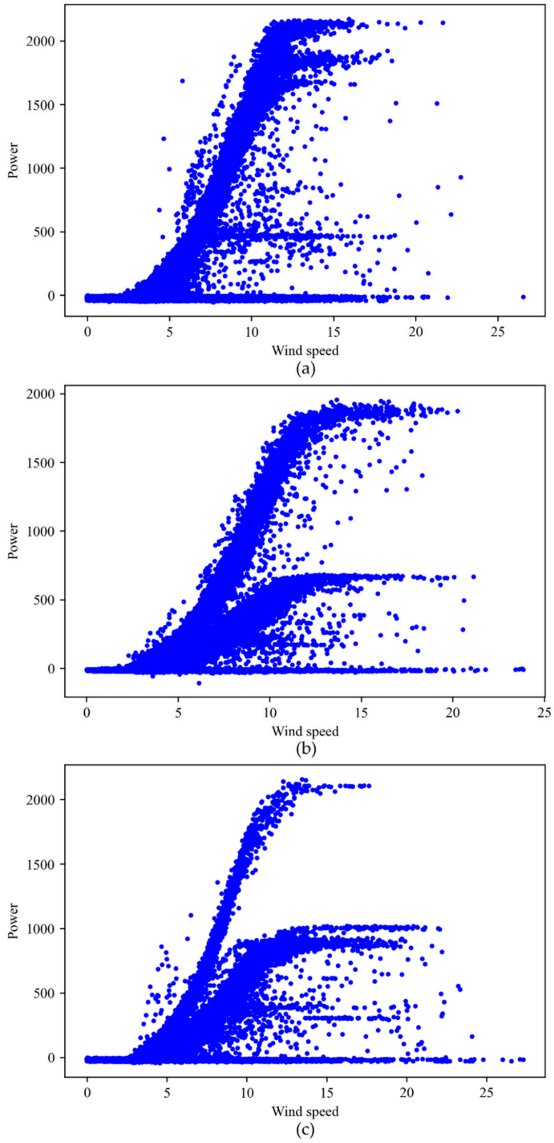

Figure 5. The original data wind speed-power scatter diagram of the three wind turbines: (**a**) #1 wind turbine; (**b**) #2 wind turbine; (**c**) #3 wind turbine.

4.2. Algorithm Experiment Process

According to Formula (2) wind speed-power characteristic, the data are preprocessed. The following types of data that do not conform to the wind speed-power characteristics can be regarded as abnormal data: (1) data with wind speed less than zero; (2) data with power not zero when the wind speed is not within the range of cut-in wind speed and

cut-out wind speed; (3) when the wind speed is within the range of cut-in wind speed and cut-out wind speed, the power is zero. Select the data part of the wind turbine that is normally connected to the grid for subsequent data cleaning.

4.2.1. Quartile Preprocessing

According to the quartile method in Section 3.1, for the three wind turbines, the wind speed interval is divided at 0.5 m/s intervals from the cut in wind speed to the cut in wind speed, and the power in each wind speed interval adopts quartile method, data with too low power in the interval is marked as abnormal data. From zero to the highest power, the power intervals are divided at intervals of 1.25% of the rated power, the wind speed in each power interval adopts the quartile method, and the data with excessive wind speed are marked as abnormal data.

Because the quartile method depends on the statistical characteristics of the data, when lots of centrally distributed abnormal data change the statistical characteristics of the original data, normal data may be mistaken for abnormal data. The quartile method is used for the three wind turbines as shown in Figure 6.

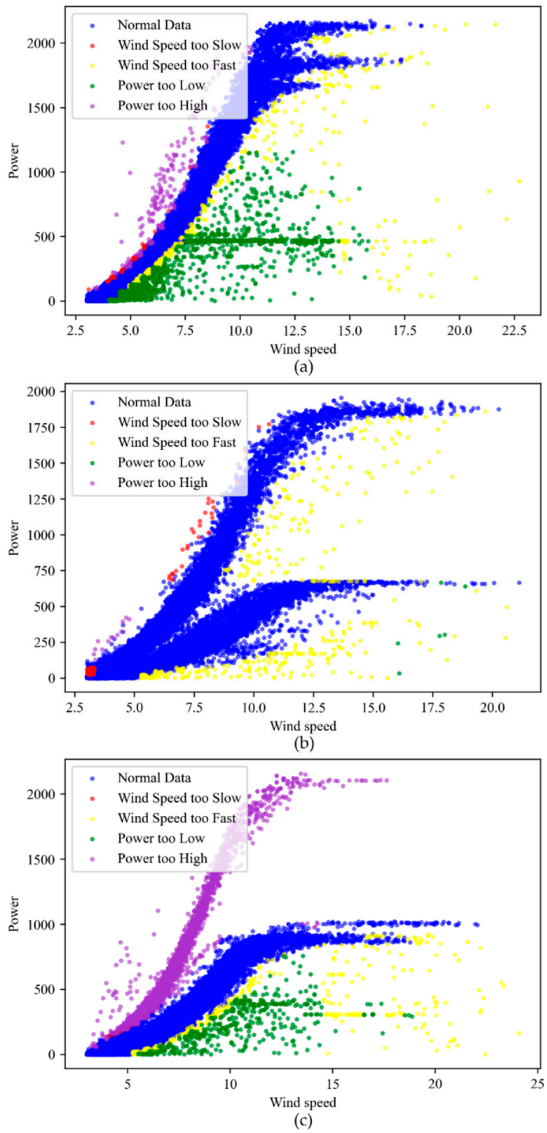

Figure 6. Three wind turbines treated by quartile method: (**a**) #1 wind turbine; (**b**) #2 wind turbine; (**c**) #3 wind turbine.

In Figure 6, the proportion of abnormal data of #1 wind turbine is less, so the quartile method can clean out abnormal data very well. However, in the #2 and #3 wind turbines, because the abnormal data of the #2 wind turbine have the same proportion as the normal data, the quartile method cannot identify the abnormal data; #3 wind turbine has a large amount of limited power abnormal data, and the proportion of abnormal data is large, so the normal data are mistakenly marked as power too high (purple in Figure 6c), and the limited power abnormal data are mistakenly marked as normal data. Because a large number of upper abnormal data and limited power abnormal data will cause marking errors, only the data with power too low in each wind speed interval and the data with wind speed too high in each power interval (green and yellow in the Figure 6) are removed by the quartile method in this method.

4.2.2. Constrained Wind Speed-Power Curve Fitting

For the #1, #2, and #3 wind turbines, according to the Formula (6) in Section 3.2, set the upper limit of the partial derivative $\alpha = 150$, the constraint margin $\delta = 0.001$, and construct the constrained optimization problem. According to Formula (7) in Section 3.3, set the initial parameter $\vec{x} = (0, 0, 0, 0)$, the initial exterior penalty factor $M^{(1)} = 1$, the amplification factor $c = 8$, the precision $\varepsilon_1 = \varepsilon_2 = 0.001$, the penalty factor control factor $R = 1000$, and the exterior penalty function method is applied to solve the optimal parameter $\vec{x}^*$. The optimal parameters $\vec{x}^*$ are substituted into the constraint functions $g_1\left(\vec{x}^*\right)$ and $g_2\left(\vec{x}^*\right)$ as shown in Table 2. The fitting curves of the #1, #2, and #3 wind turbines are shown in Figure 7.

Table 2. The result of solving for the optimal parameters.

Wind Number	Optimal Parameters $\vec{x}^*$	Constraint Function $g_1(\vec{x}^*)$	Constraint Function $g_2(\vec{x}^*)$
1	$(1.108 \times 10^{+00}, 5.222 \times 10^{-04}, 6.147 \times 10^{-01}, 2.286 \times 10^{+00})$	$2.514 \times 10^{+01}$	$1.749 \times 10^{+02}$
2	$(6.742 \times 10^{-01}, 2.833 \times 10^{-04}, 3.915 \times 10^{-01}, 3.530 \times 10^{+00})$	2.116×10^{-02}	$1.999 \times 10^{+02}$
3	$(1.256 \times 10^{+00}, 5.930 \times 10^{-04}, 7.648 \times 10^{-01}, -1.476 \times 10^{-01})$	7.126×10^{-03}	$1.999 \times 10^{+02}$

4.2.3. Improved 3-σ Division of Abnormal Data

According to Formula (9) and the values of constraint functions $g_1\left(\vec{x}^*\right)$ and $g_2\left(\vec{x}^*\right)$ in Table 2, the upper and lower limits of power are obtained for each data point in each wind speed interval. Those that are not within the upper and lower limits can be regarded as abnormal data. In Figure 7, the fitting curves of #2 and #3 wind turbines cannot fully fit the normal data even under the constrained condition due to the influence of abnormal data. However, according to Table 2, the $g_1\left(\vec{x}^*\right)$ values of the #2 and #3 wind turbines are very small, so the lower limit of the power is very close to the fitted curve, and the normal data and abnormal data can still be well divided. The results of the improved 3-σ division of abnormal data are shown in Figure 8.

Observe the result of the improved 3-σ division of abnormal data in Figure 8. For the #2 wind turbine, the result of the improved 3-σ division of abnormal data results in the truncation of the wind speed. Parts of data below the fitted curve with power within the rated power error are considered normal data. The power P_{max} corresponding to the maximum wind speed v_{max} in the normal data after the improved 3-σ division can be considered as the actual rated power. For any data point (v_i, P_i) in the data set, if $v_{max} < v_i < v_{\text{out}}$ and $0.95P_{max} < P_i < 1.05P_{max}$, then the data point (v_i, P_i) can still be considered as normal data. The final abnormal data cleaning result is shown in Figure 9.

Under the condition that other parameters remain unchanged, the value of the upper limit of the partial derivative α in Formula (6) ranges from 50 to 300, and each step is 50. The result of quadratic fitting of the cleaned data is shown in Figure 10. It can be seen from Figure 10 that the curve can be well fitted under different parameter α values, which shows that the method is not parameter-sensitive.

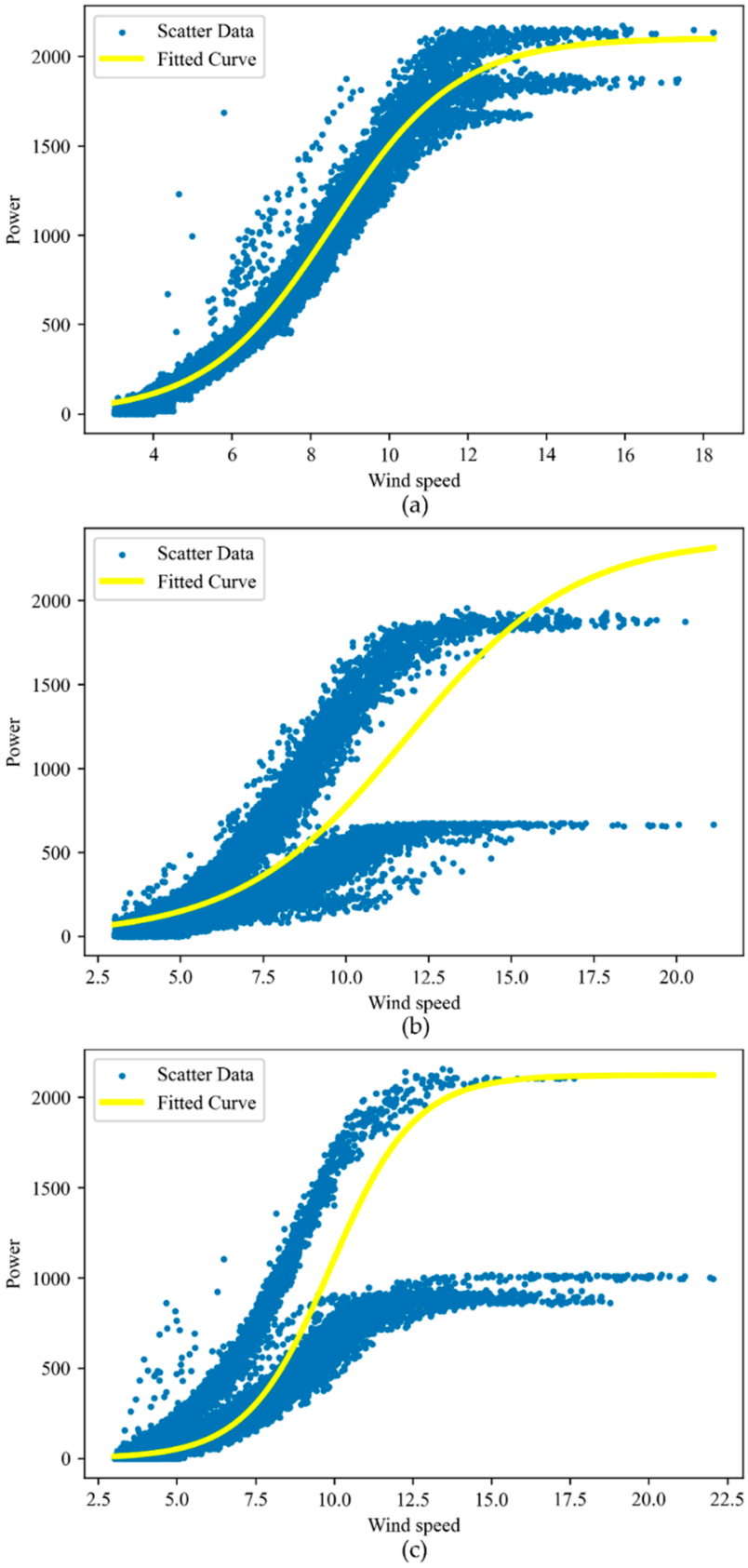

Figure 7. Curve fitting results: (**a**) #1 wind turbine; (**b**) #2 wind turbine; (**c**) #3 wind turbine.

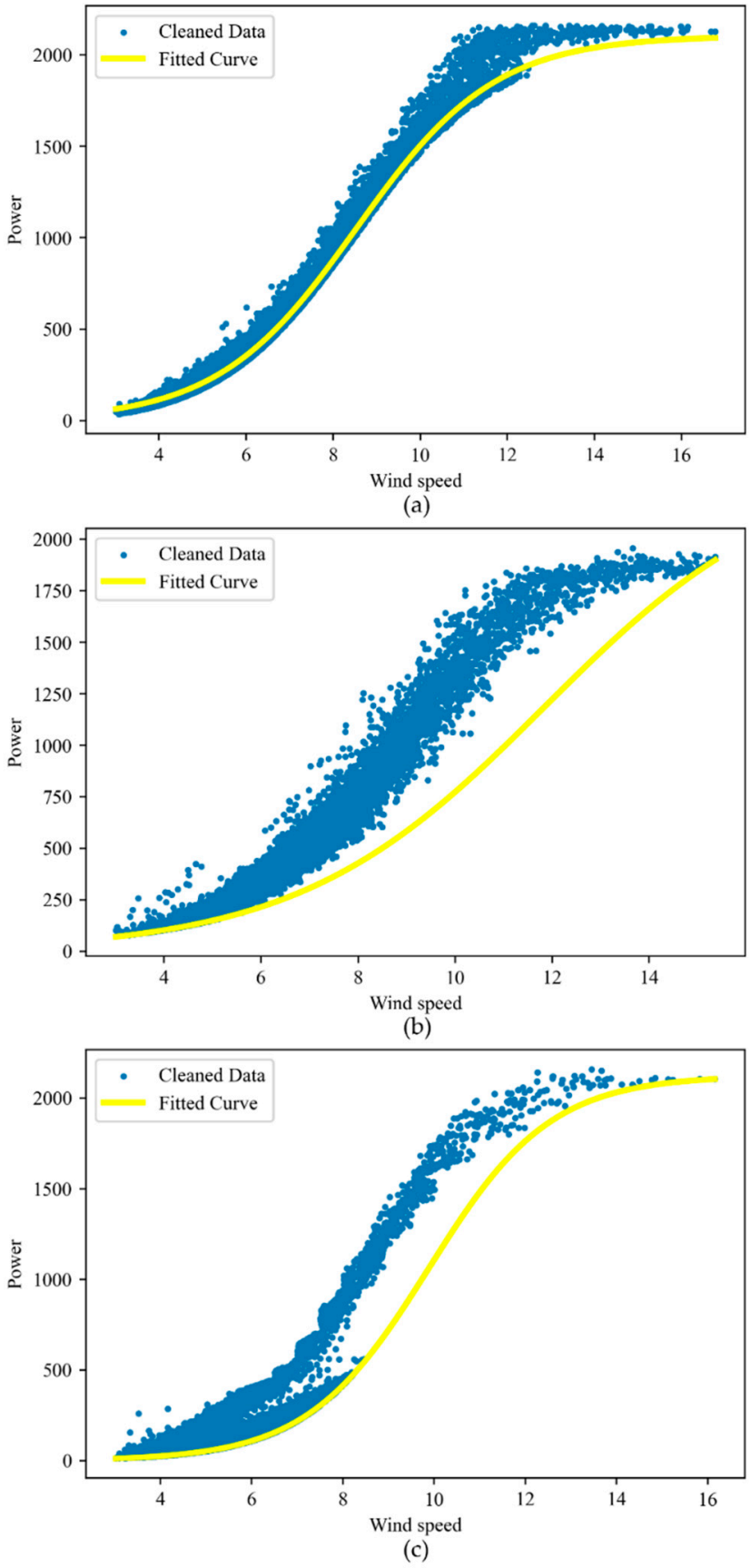

Figure 8. (**a**) #1 wind turbine; (**b**) #2 wind turbine; (**c**) #3 wind turbine.

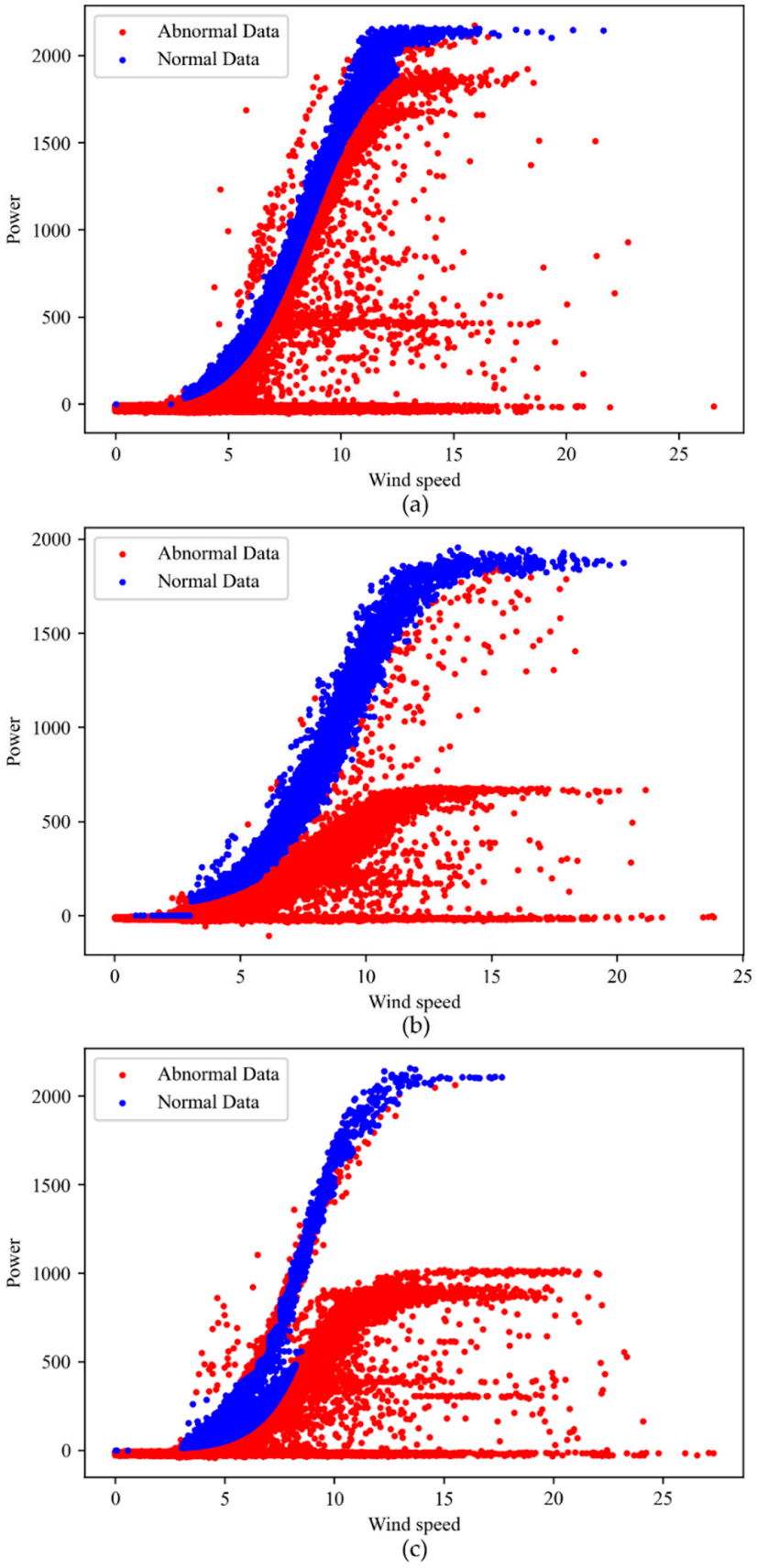

Figure 9. Abnormal data cleaning method for wind turbines based on constrained curve fitting: (**a**) #1 wind turbine; (**b**) #2 wind turbine; (**c**) #3 wind turbine.

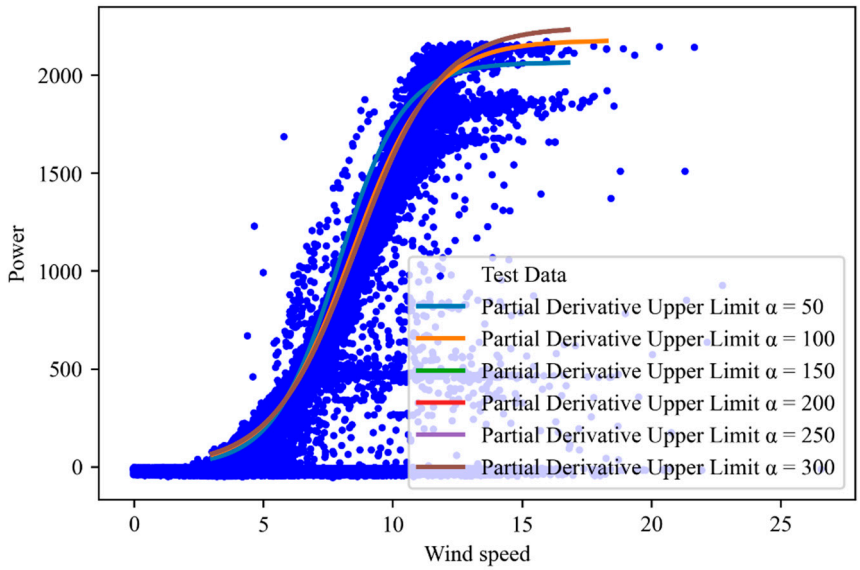

Figure 10. The results of quadratic fitting of the cleaned data under different partial derivative upper limit α values.

4.3. Algorithm Comparison

The results obtained by the algorithm in this paper are compared with the Optional Interclass Variance algorithm (OIV) [27], the Cloud Segment Optimal Entropy algorithm (CSOE) [23], and the Density-Based Spatial Clustering of Applications with Noise algorithm (DBSCAN) [20]. The comparison results are shown in Figures 11–13.

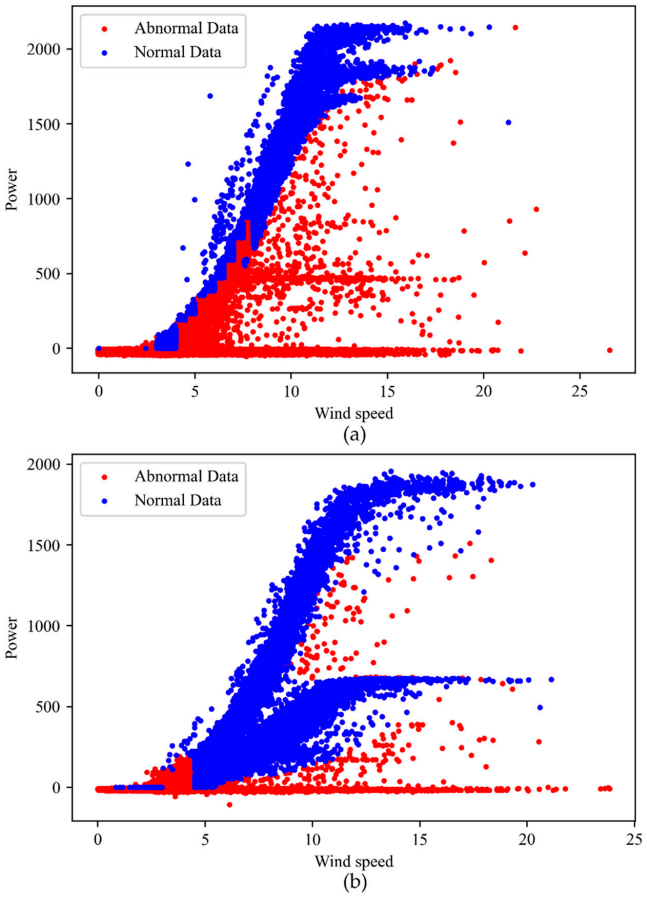

Figure 11. *Cont.*

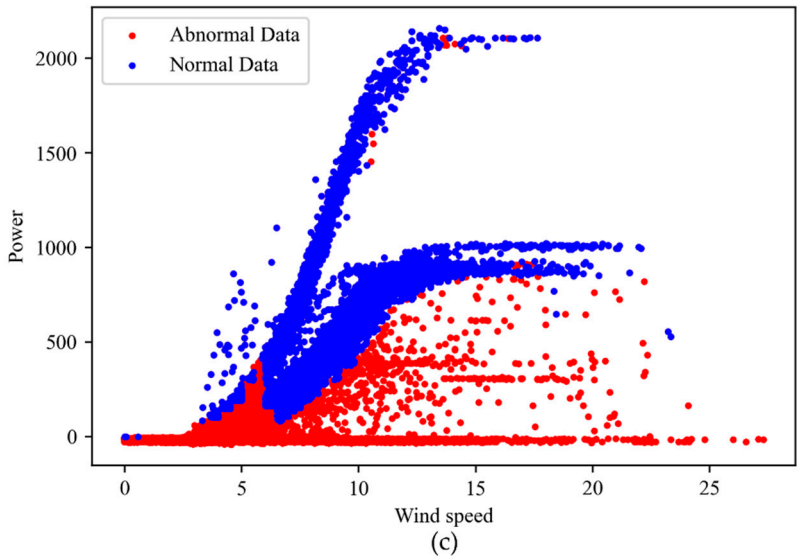

Figure 11. Abnormal data cleaning results using the OIV algorithm: (**a**) #1 wind turbine; (**b**) #2 wind turbine; (**c**) #3 wind turbine.

The results of the three methods are compared as follows: (1) The OIV algorithm needs to select different variance thresholds S according to the slip curve for different wind speed intervals. When $S = 100$, because the #1 wind turbine has a small amount of limited power abnormal data, the abnormal data cleaning is good; but when there is a large amount of limited power abnormal data such as #2 and #3 wind turbines, the effect is not good, and the upper abnormal data is not processed. (2) The CSOE algorithm also needs to select different thresholds R and r according to the change of the entropy set curve for different wind speed intervals. The CSOE algorithm processing #1 wind turbine is not as effective as the OIV algorithm and contains a large number of scattered points that have not been removed; however, the processing effect of #2 and #3 wind turbines is better than the OIV algorithm, but there are still some limited power abnormal data that have not been cleaned, and still failed to clean the upper abnormal data. (3) The DBSCAN algorithm needs to select the clustering radius eps and the minimum number of samples $MinPts$ for each wind speed interval. Take 0.3 m/s to divide the wind speed interval, when $eps = 0.02$ and $MinPts = 40$, the DBSCAN algorithm has the best processing effect on the #1 wind turbine, and can clean upper abnormal data and limited power abnormal data at the same time; but for #2 and #3 wind turbines, it does not work very well with large amounts of anomalous data changing the data density. It is found that the three methods all have the problem of difficult parameter selection, and it is not easy to realize automation. For #2 and #3 wind turbines with a large amount of limited power abnormal data, all three methods cannot clean abnormal data very well.

Compared with Figure 9, the abnormal data cleaning method for wind turbines based on constrained curve fitting can effectively clean the abnormal data of #1, #2, and #3 wind turbines at the same time, and the cleaned normal data conform to the wind speed-power characteristics and the ideal wind speed-power curve; it can process upper abnormal data and limited power abnormal data at the same time, and it can also run well in the presence of a large number of abnormal data; the parameter selection is fixed; and it is easy to realize automation.

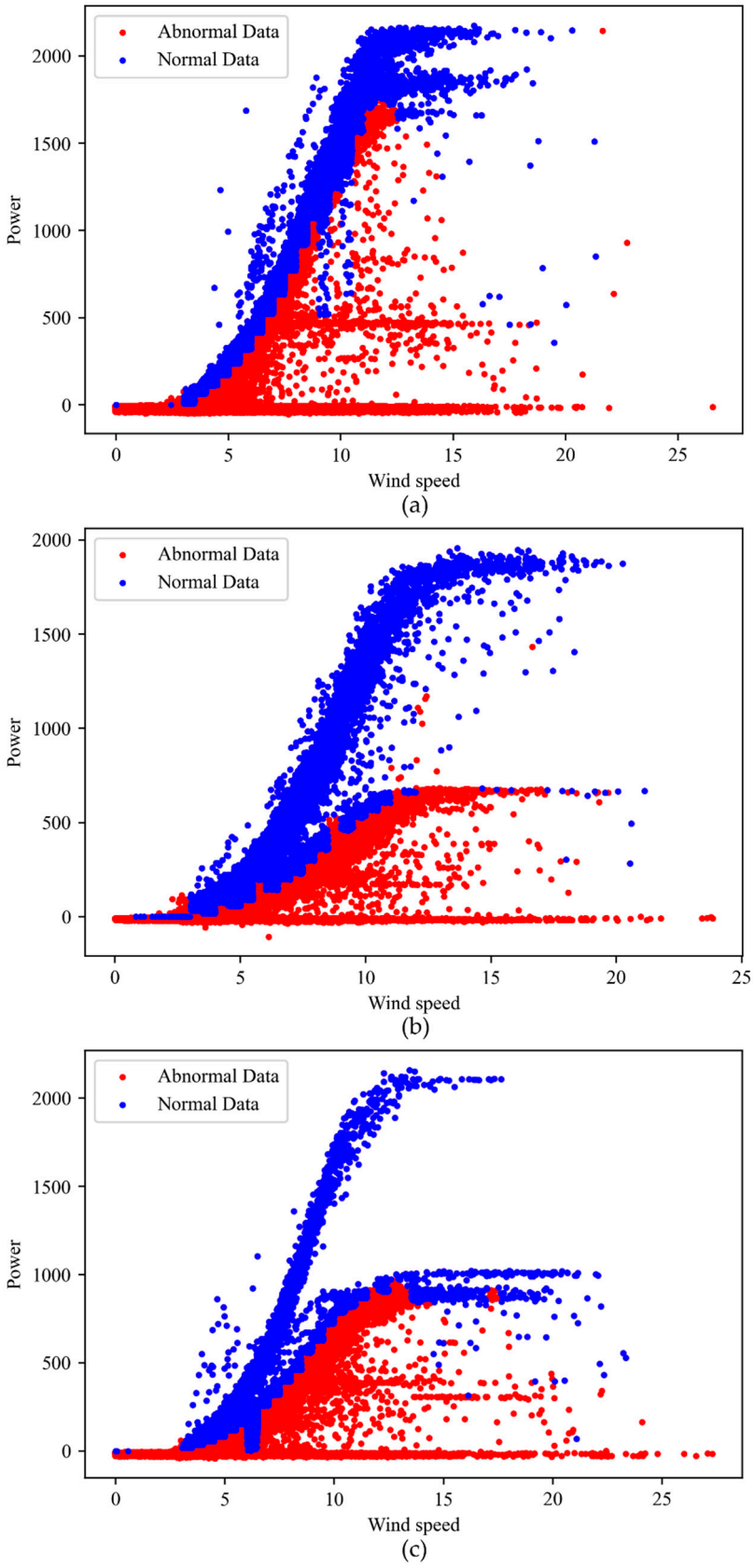

Figure 12. Abnormal data cleaning results using the CSOE algorithm: (**a**) #1 wind turbine: (**b**) #2 wind turbine; (**c**) #3 wind turbine.

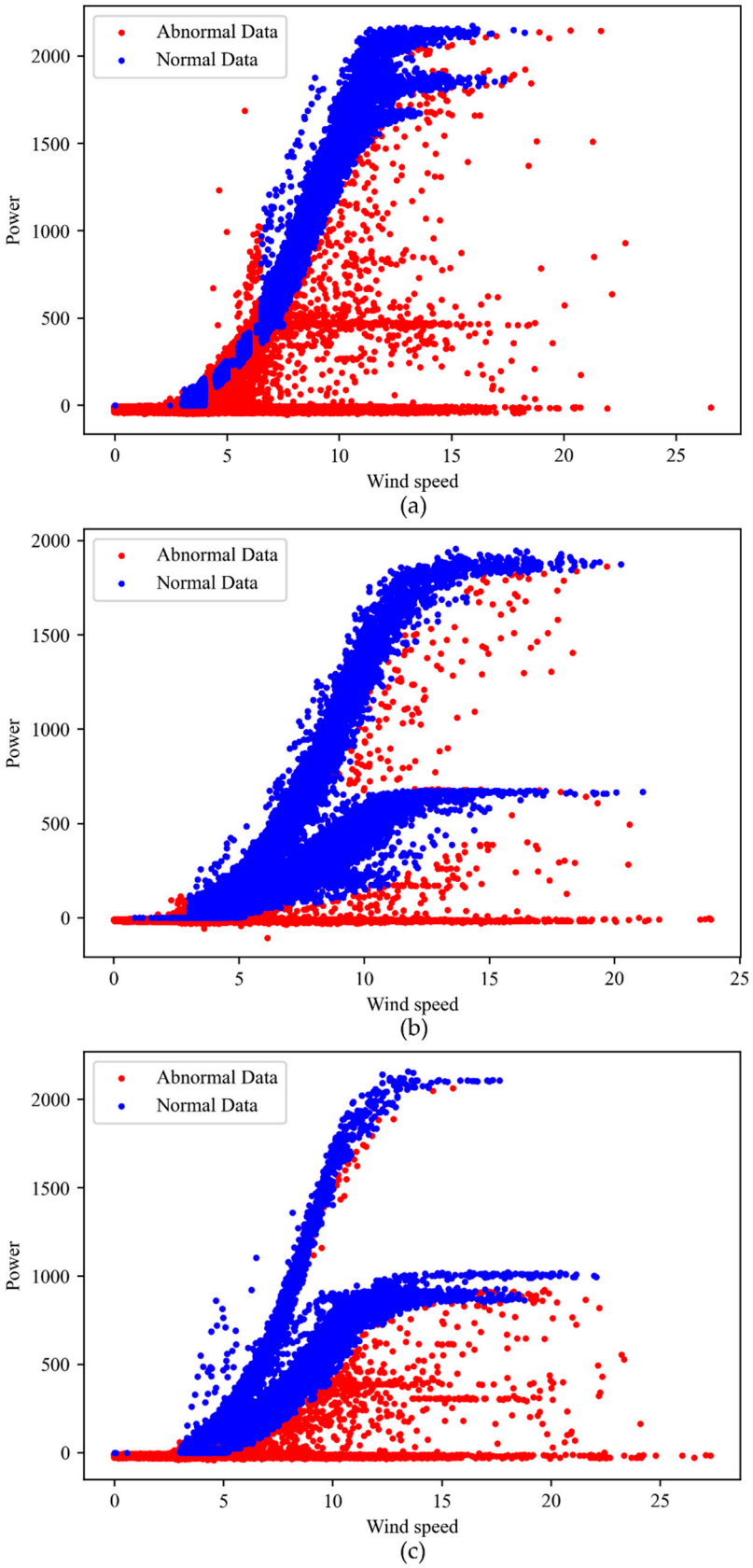

Figure 13. Abnormal data cleaning results using the DBSCAN algorithm: (**a**) #1 wind turbine; (**b**) #2 wind turbine; (**c**) #3 wind turbine.

The time taken by the algorithm and the average of entropy and hyper entropy are of great significance for practical applications, where the time taken by the algorithm evaluates the computational effort of the algorithm. Entropy is used to measure the uncertainty of qualitative concepts, which is determined by the randomness and fuzziness of concepts. Hyper entropy is used to measure the uncertainty of entropy, that is, the entropy of entropy, which is determined by the randomness and ambiguity of entropy [3]. In any wind speed interval U_v^i, the entropy En, and hyper entropy He of power can be expressed as:

$$E = \frac{\sum_{j=1}^{N} P_i^j}{N}$$

$$c_2 = \frac{\sum_{j=1}^{N} \left(P_i^j - E\right)^2}{N-1}$$

$$c_4 = \frac{\sum_{j=1}^{N} \left(P_i^j - E\right)^4}{N-1}$$

$$En = \sqrt[4]{\frac{9{c_2}^2 - c_4}{6}} \tag{10}$$

$$He = \sqrt{c_2 - \sqrt{\frac{9{c_2}^2 - c_4}{6}}} \tag{11}$$

Among them, N is the total amount of data in the wind speed interval U_v^i; P_i^j is the power value of the jth data in the wind speed interval U_v^i; E is the mean value of the power in the wind speed interval; c_2 is the second-order central moment of the power; and c_4 is the fourth-order central moment of power.

The average entropy is the average value of the entropy of the power in each wind speed interval of the cleaned data. In the wind speed interval, it reflects not only the degree of dispersion of the power but also the value range of the power, which can be applied to evaluate the stability of the segmented sequence data. The larger the average entropy and the average hyper entropy, the greater the fluctuation of power in each wind speed interval of the cleaned data, and the less thorough the cleaning.

Table 3 records the time used for data cleaning by applying the abnormal data cleaning method for wind turbines based on the Constrained Curve Fitting (CCF) algorithm, the OIV algorithm, the CSOE algorithm, and the DBSCAN algorithm to the three wind turbines, respectively, and the average entropy and the average hyper entropy of the cleaned data.

Table 3. Algorithm comparison.

Algorithm Name	Wind Number	Time (s)	The Average Entropy	The Average Hyper Entropy
CCF	#1	1.29	46.73	6.27
	#2	1.73	50.99	12.24
	#3	1.53	63.71	20.18
OIV	#1	3.15	111.93	38.03
	#2	3.14	375.16	45.21
	#3	3.35	171.34	130.37
CSOE	#1	35.52	322.64	92.32
	#2	34.62	210.87	50.16
	#3	40.21	259.60	125.48
DBSCAN	#1	0.27	99.34	31.49
	#2	0.25	381.05	23.92
	#3	0.27	171.20	124.40

It can be seen from Table 3 that the algorithm in this paper processes each wind turbine within 2 s, which meets the actual industrial needs; and the average entropy and the average hyper entropy of the three wind turbines are far lower than other algorithms. It shows that the power fluctuation in each wind speed interval of the cleaned data is small and the power is stable. This conclusion is consistent with that shown in Figures 9 and 11–13.

To sum up, the CCF algorithm has the following advantages over other traditional methods: (1) The cleaned data are more in line with the wind speed-power characteristics and the ideal wind speed-power curve; (2) the parameters are easy to select, have high robustness, and are easy to satisfy automation requirements; (3) the algorithm has a small amount of calculation and meets the needs of practical applications; (4) it can process the upper abnormal data and the limited power abnormal data at the same time, and can still work efficiently in the case of a large amount of abnormal data.

5. Conclusions

There are a lot of abnormal data in wind turbines and the abnormal data distribution of different wind turbines is different. According to the wind speed-power characteristics, the abnormal data cleaning method for wind turbines based on constrained curve fitting is proposed. Furthermore, the metric entropy and hyper entropy are used to evaluate the stability of the cleaned data, and the feasibility of the proposed model is verified experimentally.

(1) Compared with the traditional data cleaning method, the wind turbine abnormal data cleaning method based on constrained curve fitting has the advantages of insensitivity to parameters and less computation. Experiments show that the method can still perform data cleaning well in the presence of a large number of abnormal data. Compared with the traditional data cleaning method, the cleaned data are more in line with the wind speed-power characteristics and the ideal wind speed-power curve.
(2) We use entropy and hyper entropy to evaluate the stability of the cleaned data. The rationality of the index is verified by comparing the index with the data scatterplots of three wind turbines in a wind farm in eastern China after cleaning with different algorithms.
(3) Experiments show that the abnormal data cleaning method for wind turbines based on constrained curve fitting can effectively clean the data and improve the data quality, which is of great significance to the follow-up research. The next step should focus on further improving the running speed of the algorithm, improving the fitting degree of the cleaned data and the ideal wind speed-power curve, and further improving the operating efficiency.

Author Contributions: Investigation, X.Y.; Project administration, X.Y.; Software, L.Y.; Writing—original draft, Y.L.; Writing—review & editing, W.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by "The Fundamental Research Funds for the Central Universities" grant number "2022YQJD18"; "CUMTB undergraduate education and teaching reform and research funding project" grant number "J210404"; "CUMTB Postgraduate Course Ideological and political construction project 2022" grant number "YKCSZ2022004035". And The APC was funded by "The Fundamental Research Funds for the Central Universities" and "CUMTB undergraduate education and teaching reform and research funding project" and "CUMTB Postgraduate Course Ideological and political construction project 2022".

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: This work is supported by The Fundamental Research Funds for the Central Universities 2022YQJD18; CUMTB undergraduate education and teaching reform and research

funding project J210404; and CUMTB Postgraduate Course Ideological and political construction project 2022 YKCSZ2022004035.

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

References

1. Kumar, J.; Agarwal, A.; Singh, N. Design, operation and control of a vast DC microgrid for integration of renewable energy sources. *Renew. Energy Focus* **2020**, *34*, 17–36. [CrossRef]
2. Neshat, M.; Nezhad, M.M.; Abbasnejad, E.; Mirjalili, S.; Groppi, D.; Heydari, A.; Tjernberg, L.B.; Garcia, D.A.; Alexander, B.; Shi, Q.; et al. Wind turbine power output prediction using a new hybrid neuro-evolutionary method. *Energy* **2021**, *229*, 120617. [CrossRef]
3. Lin, Z.; Liu, X. Wind power forecasting of an offshore wind turbine based on high-frequency SCADA data and deep learning neural network. *Energy* **2020**, *201*, 117693. [CrossRef]
4. Tautz-Weinert, J.; Watson, S.J. Using SCADA data for wind turbine condition monitoring–a review. *IET Renew. Power Gener.* **2017**, *11*, 382–394. [CrossRef]
5. Black, I.M.; Richmond, M.; Kolios, A. Condition monitoring systems: A systematic literature review on machine-learning methods improving offshore-wind turbine operational management. *Int. J. Sustain. Energy* **2021**, *40*, 923–946. [CrossRef]
6. Zhang, J.; Jiang, N.; Li, H.; Li, N. Online health assessment of wind turbine based on operational condition recognition. *Trans. Inst. Meas. Control* **2019**, *41*, 2970–2981. [CrossRef]
7. McKinnon, C.; Turnbull, A.; Koukoura, S.; Carroll, J.; McDonald, A. Effect of time history on normal behaviour modelling using SCADA data to predict wind turbine failures. *Energies* **2020**, *13*, 4745. [CrossRef]
8. Udo, W.; Muhammad, Y. Data-driven predictive maintenance of wind turbine based on SCADA data. *IEEE Access* **2021**, *9*, 162370–162388. [CrossRef]
9. Leahy, K.; Gallagher, C.; O'Donovan, P.; O'Sullivan, D.T. Issues with data quality for wind turbine condition monitoring and reliability analyses. *Energies* **2019**, *12*, 201. [CrossRef]
10. Astolfi, D.; Castellani, F.; Natili, F. Wind turbine multivariate power modeling techniques for control and monitoring purposes. *J. Dyn. Syst. Meas. Control* **2021**, *143*, 034501. [CrossRef]
11. Long, H.; Xu, S.; Gu, W. An abnormal wind turbine data cleaning algorithm based on color space conversion and image feature detection. *Appl. Energy* **2022**, *311*, 118594. [CrossRef]
12. Su, Y.; Chen, F.; Liang, G.; Wu, X.; Gan, Y. Wind Power Curve Data Cleaning Algorithm via Image Thresholding. In Proceedings of the 2019 IEEE International Conference on Robotics and Biomimetics (ROBIO), Dali, China, 6–8 December 2019; pp. 1198–1203.
13. Liang, G.; Su, Y.; Chen, F.; Long, H.; Song, Z.; Gan, Y. Wind power curve data cleaning by image thresholding based on class uncertainty and shape dissimilarity. *IEEE Trans. Sustain. Energy* **2020**, *12*, 1383–1393. [CrossRef]
14. Wang, Z.; Wang, L.; Huang, C. A fast abnormal data cleaning algorithm for performance evaluation of wind turbine. *IEEE Trans. Instrum. Meas.* **2020**, *70*, 1–12. [CrossRef]
15. Xu, K.; Yan, J.; Zhang, H.; Zhang, H.; Han, S.; Liu, Y. Quantile based probabilistic wind turbine power curve model. *Appl. Energy* **2021**, *296*, 116913. [CrossRef]
16. Han, S.; Qiao, Y.; Yan, P.; Yan, J.; Liu, Y.; Li, L. Wind turbine power curve modeling based on interval extreme probability density for the integration of renewable energies and electric vehicles. *Renew. Energy* **2020**, *157*, 190–203. [CrossRef]
17. Seo, S.; Oh, S.I.; Kwak, H.-Y. Wind turbine power curve modeling using maximum likelihood estimation method. *Renew. Energy* **2019**, *136*, 1164–1169. [CrossRef]
18. Li, T.; Liu, X.; Lin, Z.; Morrison, R. Ensemble offshore Wind Turbine Power Curve modelling—An integration of Isolation Forest, fast Radial Basis Function Neural Network, and metaheuristic algorithm. *Energy* **2022**, *239*, 122340. [CrossRef]
19. Park, J.; Lee, J.; Oh, K.; Lee, J. Development of a Novel Power Curve Monitoring Method for Wind Turbines and Its Field Tests. *IEEE Trans. Energy Convers.* **2014**, *29*, 119–128. [CrossRef]
20. Zhao, Y.; Ye, L.; Wang, W.; Sun, H.; Ju, Y.; Tang, Y. Data-Driven Correction Approach to Refine Power Curve of Wind Farm Under Wind Curtailment. *IEEE Trans. Sustain. Energy* **2018**, *9*, 95–105. [CrossRef]
21. Yang, M.; Zhai, G.; Su, X. An Algorithm for Abnormal Data Identification of Wind Turbine Based on Wind Characteristic Analysis. In Proceedings of the 2nd World Congress on Civil, Structural, and Environmental Engineering, Barcelona, Spain, 2–4 April 2017; Volume 37, pp. 144–151. [CrossRef]
22. Xiang, L.; Deng, Z.; Zhao, Y. Anomaly Recognition Method for Wind Turbines Based on SCADA Data. *Acta Energy Sol. Sin.* **2020**, *41*, 278–284.
23. Yang, M.; Yang, Q. The Identification Research of the Wind Turbine Abnormal Data Based on the Cloud Segment Optimal Entropy Algorithm. In Proceedings of the 3rd World Congress on Civil, Structural, and Environmental Engineering, Budapest, Hungary, 8–10 April 2018; Volume 38, pp. 2294–2301+2539. [CrossRef]
24. Wang, X.; Wang, Z. Wind speed-power data cleaning of wind turbine based on improved bin algorithm. *Chin. J. Intell. Sci. Technol.* **2020**, *2*, 62–71.

25. Han, B.; Xie, H.; Shan, Y.; Liu, R.; Cao, S. Characteristic Curve Fitting Method of Wind Speed and Wind Turbine Output Based on Abnormal Data Cleaning. In Proceedings of the 2021 International Conference on Advanced Technologies and Applications of Modern Industry (ATAMI 2021), Wuhan, China, 19–21 November 2021; Volume 2185, p. 012085.
26. Zou, T.; Gao, Y.; Yi, H.; Xu, C.; Xia, R.; Wu, C. Processing of Wind Power Abnormal Data Based on Thompson tau-quartile and Multi-point Interpolation. *Autom. Electr. Power Syst.* **2020**, *44*, 156–162.
27. Lou, J.; Xu, J.; Lu, H.; Qu, Z.; Li, S.; Liu, R. Wind Turbine Data-cleaning Algorithm Based on Power Curve. *Autom. Electr. Power Syst.* **2016**, *40*, 116–121.
28. Wang, S.; Zhang, Z.; Wang, P.; Tian, Y. Failure warning of gearbox for wind turbine based on 3σ-median criterion and NSET. *Energy Rep.* **2021**, *7*, 1182–1197. [CrossRef]
29. Tao, L.; Siqi, Q.; Zhang, Y.; Shi, H. Abnormal detection of wind turbine based on SCADA data mining. *Math. Probl. Eng.* **2019**, *2019*, 5976843. [CrossRef]
30. Luo, Z.; Fang, C.; Liu, C.; Liu, S. Method for Cleaning Abnormal Data of Wind Turbine Power Curve Based on Density Clustering and Boundary Extraction. *IEEE Trans. Sustain. Energy* **2021**, *13*, 1147–1159. [CrossRef]
31. Wang, Y.; Liu, H.; Song, P.; Hu, Z.; Deng, X.; Wu, L. An approach for the cleaning of abnormal wind turbine operation data based on multi-phase progressive recognition. *Renew. Energy Resour.* **2020**, *38*, 1470–1476. [CrossRef]
32. Trivellato, F.; Battisti, L.; Miori, G. The ideal power curve of small wind turbines from field data. *J. Wind. Eng. Ind. Aerodyn.* **2012**, *107–108*, 263–273. [CrossRef]
33. Si, C.; Lan, T.; Hu, J.; Wang, L.; Wu, Q. Penalty parameter of the penalty function method. *Control Decis.* **2014**, *29*, 1707–1710. [CrossRef]