energies

MDPI

Article

Cleaning of Abnormal Wind Speed Power Data Based on Quartile RANSAC Regression

Fengjuan Zhang, Xiaohui Zhang *, Zhilei Xu, Keliang Dong, Zhiwei Li and Yubo Liu

College of Electrical Engineering, Henan University of Technology, Zhengzhou 450001, China; fjzhang@stu.haut.edu.cn (F.Z.); zhileixu@haut.edu.cn (Z.X.); keliangdong@haut.edu.cn (K.D.); 2023930989@stu.haut.edu.cn (Z.L.); yuboliu@stu.haut.edu.cn (Y.L.)
* Correspondence: zhangxh@haut.edu.cn

Abstract: The combined complexity of wind turbine systems and harsh operating conditions pose significant challenges to the accuracy of operational data in Supervisory Control and Data Acquisition (SCADA) systems. Improving the precision of data cleaning for high proportions of stacked abnormalities remains an urgent problem. This paper deeply analyzes the distribution characteristics of abnormal data and proposes a novel method for abnormal data cleaning based on a classification processing framework. Firstly, the first type of abnormal data is cleaned based on operational criteria; secondly, the quartile method is used to eliminate sparse abnormal data to obtain a clearer boundary line; on this basis, the Random Sample Consensus (RANSAC) algorithm is employed to eliminate stacked abnormal data; finally, the effectiveness of the proposed algorithm in cleaning abnormal data with a high proportion of stacked abnormalities is verified through case studies, and evaluation indicators are introduced through comparative experiments to quantitatively assess the cleaning effect. The research results indicate that the algorithm excels in cleaning effectiveness, efficiency, accuracy, and rationality of data deletion. The cleaning accuracy improvement is particularly significant when dealing with a high proportion of stacked anomaly data, thereby bringing significant value to wind power applications such as wind power prediction, condition assessment, and fault detection.

Keywords: data cleaning; quartile; RANSAC; wind power curve; wind turbine

Citation: Zhang, F.; Zhang, X.; Xu, Z.; Dong, K.; Li, Z.; Liu, Y. Cleaning of Abnormal Wind Speed Power Data Based on Quartile RANSAC Regression. *Energies* **2024**, *17*, 5697. https://doi.org/10.3390/en17225697

Academic Editor: Davide Astolfi

Received: 26 September 2024
Revised: 29 October 2024
Accepted: 5 November 2024
Published: 14 November 2024

1. Introduction

As the energy crisis intensifies, renewable energy generation has emerged as a crucial means to alleviate the crisis and improve environmental quality. Wind power generation, with its clean and renewable characteristics, is leading a profound transformation in the energy sector. In 2023, the global installed capacity of new wind power reached a record-breaking 117 GW, marking a year-on-year increase of 50% [1]. Furthermore, the Global Wind Energy Council (GWEC) has revised upwards its growth forecast for 2024–2030 by 10% to 1210 GW, highlighting the robust growth momentum of global wind energy. To ensure efficient supervision and maintenance management, most wind turbine systems are equipped with Supervisory Control and Data Acquisition (SCADA) systems for data collection. Among these data, the wind speed-power output curves are indispensable in many applications, serving as vital reference indicators for assessing the power generation efficiency of individual wind turbines and the overall operational status of wind farms [2]. They are frequently utilized in wind power forecasting [3–5], performance and condition monitoring of wind turbines [6–8], as well as fault diagnosis of wind turbine systems [9–11]. However, due to factors such as extreme weather conditions, wind curtailment, sensor failures, communication malfunctions, and others, the data collected by SCADA systems often contain a significant amount of abnormal data, which undermines the accuracy of wind turbine research. Consequently, effectively cleaning the abnormal values in wind power data has become a crucial and indispensable step.

Currently, domestic and foreign scholars have carried out relevant research on wind power abnormal data cleaning and achieved many results. These research results can be mainly divided into the following three categories.

(1) The first category mainly relies on statistical and clustering methods for data cleaning. Lou et al. employed the Optimal In-group Variance (OIV) method for cleaning [12], which, despite its rapid identification capability, is susceptible to the influence of data grouping methods. Zheng et al. utilized the Local Outlier Factor (LOF) to distinguish between normal and abnormal data [13], showcasing strong adaptability and the ability to handle data with different density distributions. Reference [14] initially used the quartile method to eliminate dispersed data, followed by k-means clustering to clean the stacked data, although the selection of parameters significantly impacts the clustering effectiveness. Zhao et al. proposed a strategy combining Density-Based Spatial Clustering of Applications with Noise (DBSCAN) and the quartile method [15]. In reference [16], a combined algorithm of Isolation Forest and mean shift was constructed. Although it can achieve efficient cleaning, its cleaning results in datasets with high noise levels may be affected. Luo et al. designed a method based on density clustering and boundary extraction, but this approach incurs higher time costs [17].

(2) The second category determines the upper and lower boundaries of the wind power curve for cleaning. In reference [18], a wind power curve model was established using the copula conditional function, which is highly effective for identifying sparse anomalous data. Villanueva and Feijóo proposed a real power curve model that fits the wind power within various wind speed ranges to a normal probability distribution, considering data exceeding three standard deviations as anomalies [19]. Wang et al. improved the binning algorithm for regional calculation [20], but data at regional boundaries may be difficult to accurately fit, affecting accuracy. Cleaning through neural networks allows for the adjustment of network structures and parameters to address different types of anomalies, such as using the Artificial Neural Network (ANN) algorithm [21], as well as Graph Convolutional Neural Networks combined with Long Short-Term Memory networks (GCN-LSTM) for cleaning [22]. Neural network algorithms can automatically extract features; however, when the model becomes too complex, there is a risk of overfitting.

(3) The third category adopts the method of image processing for wind power curves. Liang et al. use the pixel counting method to generate feature grayscale images and combine the image threshold segmentation method to eliminate abnormal data [23]. Wang et al. proposed a fast data cleaning algorithm to maintain the longest continuous pixels in each column and each row of the binary curve image [24]. Long et al. proposed a wind power abnormal data cleaning algorithm based on color space transformation and image feature detection of wind power curve images [25]. However, image processing methods require powerful computing resources and often entail relatively high time costs. In summary, the relevant research that has been carried out provides references and theoretical support for subsequent work.

Although many scholars have conducted a lot of research on abnormal data cleaning, the cleaning technology for wind speed power abnormal data with a high proportion of stacked abnormalities still needs to be further improved, and there is still a lack of relevant research. Therefore, this paper proposes a new method for abnormal data cleaning based on a classification processing framework. This method is not subject to constraints or statistical assumptions about the data, and it directly performs unsupervised learning driven by the data. First, based on the distribution characteristics of abnormal data in wind power curves, this method preprocesses the first type of abnormal data based on operating criteria. Second, a combination of the quartile method and the Random Sample Consensus (RANSAC) algorithm is used to accurately clean the second type of sparse abnormal data and the third type of stacked abnormal data. Finally, the effectiveness of the quartile-RANSAC algorithm for cleaning data with a high proportion of abnormalities

is verified through case analysis. To further accurately evaluate the performance of this algorithm, this paper conducts a thorough comparison and verification from four core dimensions: cleaning effect, efficiency, accuracy, and rationality of data deletion. The research results show that when faced with high proportions of stacked abnormal data, the proposed method significantly optimizes the cleaning effect. This study not only effectively enhances the accuracy of operational data for wind turbine generators but also provides a new approach in the field of abnormal data cleaning, offering theoretical and technical support for ensuring smooth dispatch and safe, stable operation of wind farms.

This paper is divided into five sections. Section 2 analyzes the distribution characteristics of abnormal data in power curves and categorizes the abnormal data into three types. Section 3 presents the principle and framework of the algorithm proposed in this paper. In Section 4, the effectiveness of the algorithm proposed in this paper was verified through experiments, and it was compared with three other types of algorithms. Evaluation indicators were used to assess and discuss the results. Section 5 serves as the conclusion of this paper, summarizing the main research findings and conclusions.

2. Wind Power Curve

The wind power curve, formed by data collected by the SCADA system, serves as one of the key indicators for evaluating the performance of wind turbines. However, the directly collected SCADA data contain a large amount of abnormal data, which significantly interferes with the assessment of wind power generation performance. The factors causing abnormalities are diverse, such as extreme weather, wind curtailment, and turbine failures. Based on the distribution characteristics of the data within the wind power curve, these abnormalities can be roughly classified into three categories, and the distribution of each type of abnormal data is shown in Figure 1.

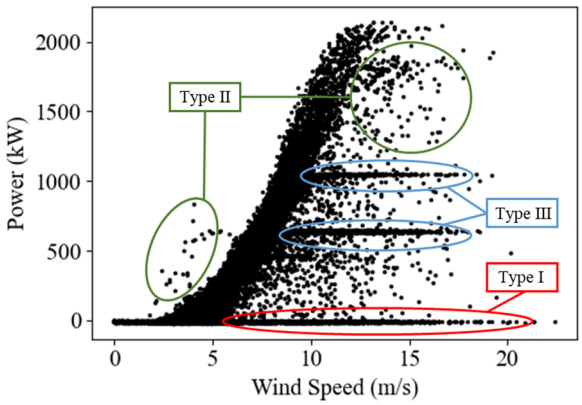

Figure 1. Classification of abnormal data in wind power curves.

(1) **Type I abnormal data are the lower horizontal band abnormal data:**
The first type of abnormal data refers to the abnormal data where the wind speed is greater than the cut-in wind speed but the power is less than or equal to zero. This type of abnormal data exhibits distinct characteristics in the wind power curve graph, forming a horizontal band region composed of dense data points at the lower end of the curve. The main factors contributing to this type of abnormality include internal failures of wind turbines and shutdowns for maintenance.

(2) **Type II abnormal data are sparse abnormal data:**
The second type of abnormal data exhibits a unique distribution pattern, appearing as scattered and irregular data points. These data points display significant randomness but maintain a certain correlation with the standard power curve. The main factors contributing to this type of abnormality include meteorological fluctuations, signal transmission noise interference, sensor failures, and various other unpredictable random factors.

(3) **Type III abnormal data are stacked abnormal data:**
The third type of abnormal data typically appears over a continuous period, clustering into one or more distinct horizontal data bands in the middle region of the power curve. The emergence of this type of abnormality is closely related to issues such as wind curtailment and communication failures. The technical factors directly leading to wind curtailment include power system failures, insufficient system frequency regulation capabilities, and inadequate transmission and storage technologies [26]. In particular, wind curtailment and power rationing have become prominent issues restricting the sustainable and healthy development of the wind power industry.

3. Building an Outlier Data Cleaning Model Based on Quartile RANSAC

An anomaly data cleaning algorithm model based on quartile RANSAC is constructed by analyzing the abnormal data distribution in the wind speed-power scatter plot of the wind turbine and aiming at the data cleaning with a high stacking anomaly proportion. Combining the quartile method with the RANSAC algorithm provides a novel method for the identification and cleaning of abnormal wind power data. The algorithm framework is shown in Figure 2.

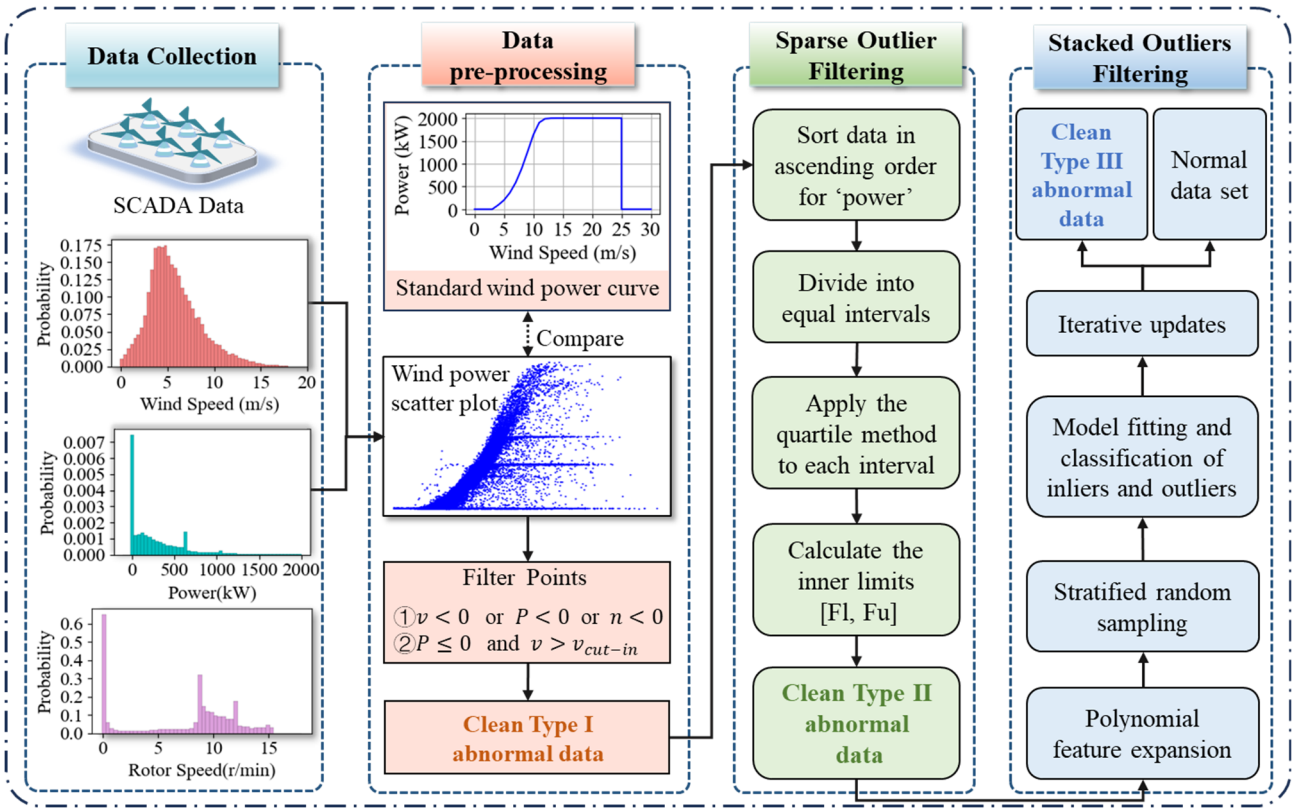

Figure 2. Framework of abnormal data cleaning algorithm based on quartile RANSAC.

The quartile RANSAC algorithm mainly includes the following steps:

1. Data collection: utilize the SCADA system to collect wind speed, power, and rotational speed data from the wind farm to form a dataset.
2. Data preprocessing: compare the wind speed-power scatter plot with the standard wind power curve and filter the first type of abnormal data based on the basic operating principles of the standard wind power curve. Eliminate data points where wind speed (v), power (P), or rotational speed (n) are less than zero, and mark data as abnormal when the power is less than or equal to zero while the wind speed is greater than the cut-in wind speed (v_{cut_in}).

3. Elimination of sparse abnormal data using the quartile method: sort the preprocessed data pairs by power in ascending order and divide the data into equal interval power bins. Apply the quartile method to filter the data within each power bin, marking wind speed data that fall outside the inner limits $[F_l, F_u]$ as abnormal and removing them from the dataset.
4. Elimination of stacked data based on RANSAC regression fitting: extend the original two-dimensional data to a three-dimensional space through polynomial features to fit more complex nonlinear relationships. On this basis, the RANSAC regression algorithm will be employed to predict wind speed values. Use random sample points to fit a model, calculate the distance of all data points to the fitted model, and classify points into inliers and outliers based on a threshold. Continuously update and iterate until the model performance is optimized. Finally, determine whether a data point exceeds the threshold; if so, it is classified as an abnormal point; otherwise, it is considered a normal point.

3.1. Data Preprocessing

Eliminate Type I abnormal data based on operational criteria. Firstly, considering practical factors, data with wind speed, power, and rotational speed less than zero need to be eliminated. Secondly, when the wind speed is greater than the cut-in wind speed, which is the minimum wind speed at which the wind turbine can generate electricity, the power should be greater than zero. Therefore, data with power less than or equal to zero are eliminated, completing the cleaning of Type I abnormal data.

3.2. Quartile Method

After data preprocessing, the quartile method is used to clean Type II abnormalities. The quartile method is a commonly used data analysis technique in statistics. Quartiles are the three values that divide an ordered data sample into four equal parts, represented by the first, second, and third quartiles, respectively, with each part containing 25% of the overall data. The schematic diagram of the quartile method is shown in Figure 3.

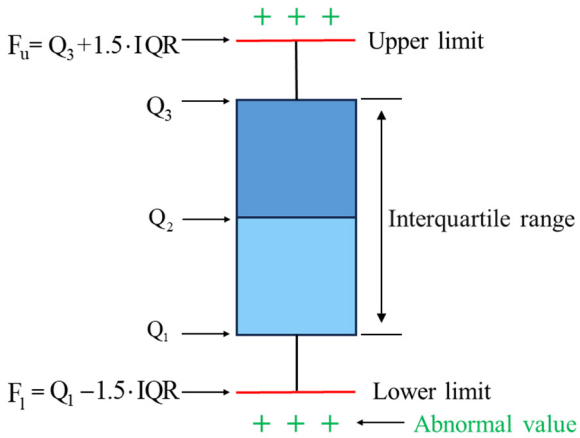

Figure 3. Schematic diagram of the quartile method.

For a sample $X = [x_1, x_2, \ldots, x_n]$ sorted in ascending order, the quartiles are calculated as follows:

1. Calculate the second quartile Q_2, which is the median:

$$Q_2 = \begin{cases} x_{\frac{n+1}{2}} & n = 2k+1; k = 0, 1, 2, \ldots \\ \frac{x_{\frac{n}{2}} + x_{\frac{n+2}{2}}}{2} & n = 2k; k = 1, 2, \ldots \end{cases} \quad (1)$$

2. Calculate the first and third quartiles Q_1 and Q_3:
When $n = 2k$ ($k = 1, 2, \ldots$), divide X into two parts from Q_2, with Q_2 excluded from

both parts of the data, and calculating the medians of the two parts, Q_2' and Q_2'', then $Q_1 = Q_2'$, $Q_3 = Q_2''$.
When $n = 4k + 3$ ($k = 0, 1, 2, \ldots$), there are

$$\begin{cases} Q_1 = 0.75x_{k+1} + 0.25x_{k+2} \\ Q_3 = 0.25x_{3k+2} + 0.75x_{3k+3} \end{cases} \quad (2)$$

When $n = 4k + 1$ ($k = 0, 1, 2, \ldots$), there are

$$\begin{cases} Q_1 = 0.25x_k + 0.75x_{k+1} \\ Q_3 = 0.75x_{3k+1} + 0.25x_{3k+2} \end{cases} \quad (3)$$

3. The interquartile range IQR can be obtained by calculating Q_1 and Q_3:

$$IQR = Q_3 - Q_1 \quad (4)$$

4. Based on the IQR, the inner limits $[F_l, F_u]$ for identifying outliers in the data sample X are determined as follows:

$$[F_l, F_u] = [Q_1 - 1.5IQR, Q_3 + 1.5IQR] \quad (5)$$

where F_l represents the lower limit and F_u represents the upper limit, data points falling outside the inner limits $[F_l, F_u]$ are considered outliers.

The quartile method is not only applicable to different data distributions but also can effectively resist the influence of extreme values in the data on statistical results, showing good robustness. In addition, this algorithm also demonstrates efficient computing capabilities in large-scale data analysis. Compared with the original data, the boundaries of abnormal data clusters are more obvious after adopting the quartile method, which enables the next step of regression analysis to achieve better results. However, this method has certain limitations and can only correctly and effectively identify abnormal data when the proportion of abnormal data is small. In the problem studied in this paper, the amount of abnormal data is comparable to that of normal data. Using the quartile method alone will not be able to effectively identify abnormal data and may even mistakenly delete normal data. Therefore, this paper only uses the quartile method to eliminate sparse outliers.

3.3. RANSAC Regression Algorithm

For Type III anomalies, the RANSAC regression algorithm is used for cleaning. The principle of the RANSAC algorithm is to robustly estimate the parameters of a mathematical model from a dataset containing a large amount of noise or outliers through an iterative approach [27]. The principle of RANSAC is illustrated in Figure 4. The algorithm first randomly selects a minimal subset of data points that satisfy predefined conditions and constructs an initial model based on this subset. Then, the algorithm evaluates and expands this subset to include more data points that are consistent with the initial model, thereby forming a larger dataset with higher data consistency. In this way, RANSAC can effectively reduce the interference of outliers or noise on model estimation and ensure the accuracy of model parameters. Finally, through multiple iterations and optimizations, the algorithm finds the optimal model parameter estimation that maximizes the utilization of valid data points (i.e., inliers) and minimizes the influence of outliers. This principle makes RANSAC perform well in processing datasets containing a large amount of noise or outliers, providing robust and reliable model parameter estimation.

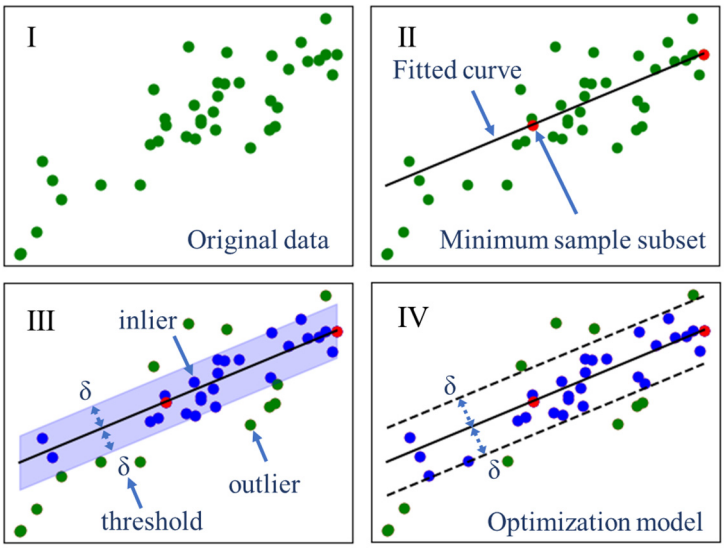

Figure 4. Principle diagram of RANSAC regression for outlier detection.

Unlike the quartile method, the RANSAC algorithm is capable of robustly fitting models even in datasets containing a significant amount of noise and outliers. By iteratively selecting the best model with the largest number of inliers, RANSAC eliminates the need to pre-specify the initial values or ranges of model parameters, thereby reducing human intervention and providing more accurate and reliable model estimates. Furthermore, the algorithm's strong adaptability and flexibility allow it to dynamically adjust and refine model parameters for optimal fitting, enabling RANSAC to often deliver excellent results when dealing with data of varying complexity.

4. Case Study

4.1. Data Description

To verify the effectiveness of the proposed wind speed and power abnormal data cleaning method, a case study is conducted using SCADA operational data from a wind farm in China. The dataset records the SCADA operational data of 12 wind turbines for one year, with data recorded every 10 min. The key technical indicators of the wind turbines are: rated power of 2000 kW, rotor diameter of 99 m, cut-in wind speed of 3 m/s, and cut-out wind speed of 25 m/s.

It is evident that the specific locations and quantities of abnormal data vary among the turbines. Therefore, this study selects wind turbines No. 2, 3, 7, and 8, which best represent typical types of abnormal data, to demonstrate the effectiveness of the proposed algorithm in identifying and eliminating various types of abnormalities. The operational data of these four wind turbines are collected by the SCADA system every 10 min, totaling 159,644 data points.

4.2. Case Study on Data Cleaning of Wind Turbine with High Proportion of Stacked Abnormalities

The SCADA data of the No. 3 wind turbine generator set are used to clean abnormal data in order to verify the effectiveness of the proposed method. The stacking abnormality proportion of the No. 3 wind turbine generator set is significantly higher than that of the other three typical types of abnormal data, which was suitable for verifying the cleaning effect of the new method proposed in this paper for cleaning abnormal data with a high stacking proportion. The cleaning results are shown in Figure 5.

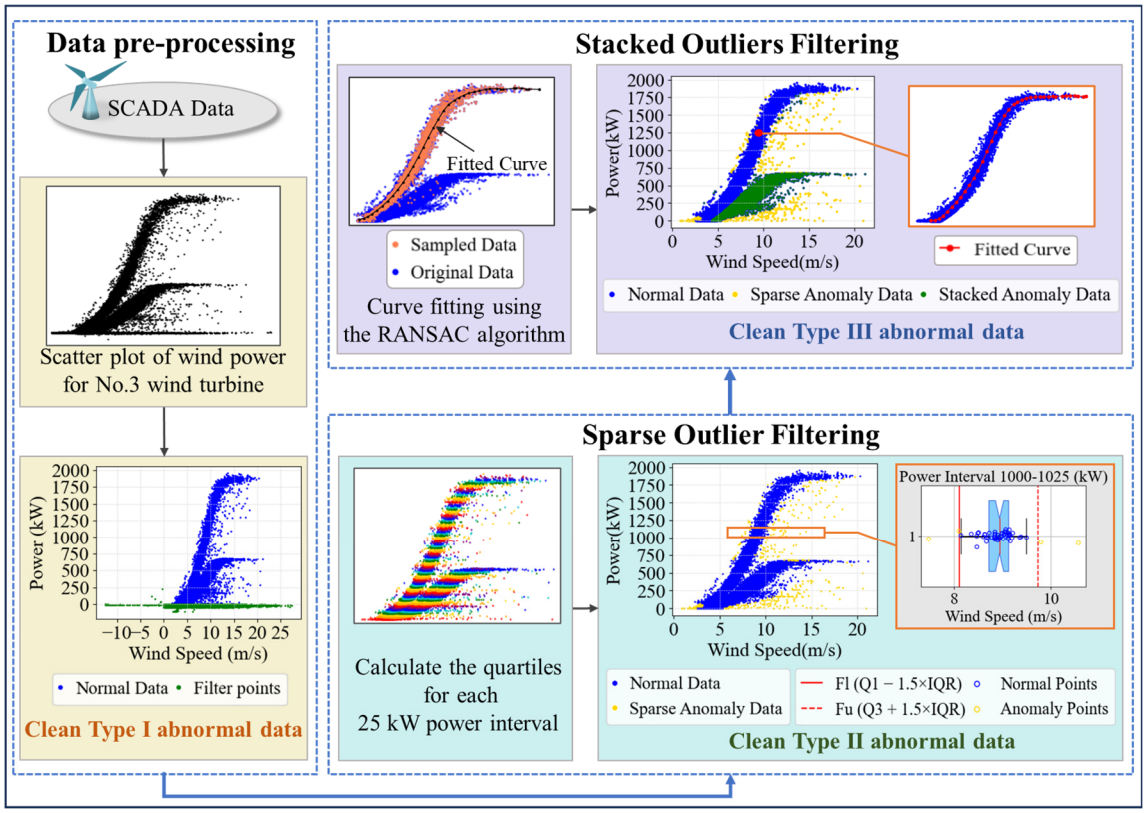

Figure 5. Cleaning process of abnormal data for No. 3 wind turbine generator set based on quartile RANSAC algorithm.

(1) Data Preprocessing
Under normal operating conditions, it is unreasonable for the generator to produce negative wind power. Therefore, values with wind speed, power, and rotational speed less than zero are eliminated from the scatter plot. Based on the basic operating principles of wind turbines, data with wind speeds exceeding the cut-in wind speed but power less than or equal to zero are also marked as abnormal, accurately identifying the first type of abnormal data.

(2) Elimination of Sparse Outliers
After preprocessing, the data are sorted in ascending order of power and divided into intervals with a spacing of 25 kW. The quartile method is then applied to each interval of data to eliminate abnormal data points lying outside the inner limits. After data preprocessing and the application of the quartile method, most sparse outliers have been eliminated, revealing a clearer data distribution profile, especially with very distinct boundaries for stacked outliers. Examining the box plot generated by the quartile method for data within the power interval [1000, 1025] kW, it can be observed that most data points within the interval fall within the inner limits $[F_l, F_u]$, identified as normal data and marked blue, while those outside the inner limits are recognized as sparse outliers and marked gold.

(3) Elimination of Stacked Outliers
When performing RANSAC regression fitting, the two-dimensional data are first extended to three-dimensional data through polynomial feature expansion to better perform nonlinear fitting. Then, hierarchical sampling of the minimum effective sample subset is performed on the data, and the corresponding model parameters are calculated using the least variance estimation method. Subsequently, the deviation between each sample data point and the estimated model is calculated. Based on this, the deviation is compared with the threshold. If the deviation is less than the threshold, it is considered normal data; if the deviation is greater than the threshold,

it is identified as abnormal data. Finally, iterations continue until the model achieves the optimal effect. As shown in Figure 5, during curve fitting, the clear boundaries of stacked outliers reduce their impact on the fitting effect. After the RANSAC regression algorithm is used to identify stacked outliers, the stacked abnormal data are clearly marked, and the wind speed-power curve is clearly outlined. Gold represents the elimination of sparse outliers, green represents the elimination of stacked outliers, and blue represents normal data. It can be seen that the wind speed-power curve is clearly identified. The power curve fitted by the "bin" method on the cleaned data is highly consistent with the standard power curve, demonstrating the effectiveness of the new method proposed in this paper for cleaning abnormal data with a high proportion of stacked outliers.

4.3. Algorithm Comparison and Analysis

4.3.1. Comparative Experiment

To evaluate the strengths and weaknesses of the quartile RANSAC algorithm, a comparative experiment was conducted with three existing algorithms: Quartile, Isolation Forest, and k-means. The cleaning effectiveness of each algorithm was measured by visualizing the post-cleaning wind speed-power curves. A high degree of consistency between the cleaned data and the standard power curve indicates effective cleaning, while significant deviations suggest ineffectiveness. The results of the comparative experiments are presented in Figures 6–9. Among them, green represents abnormal data, while normal data is marked as blue.

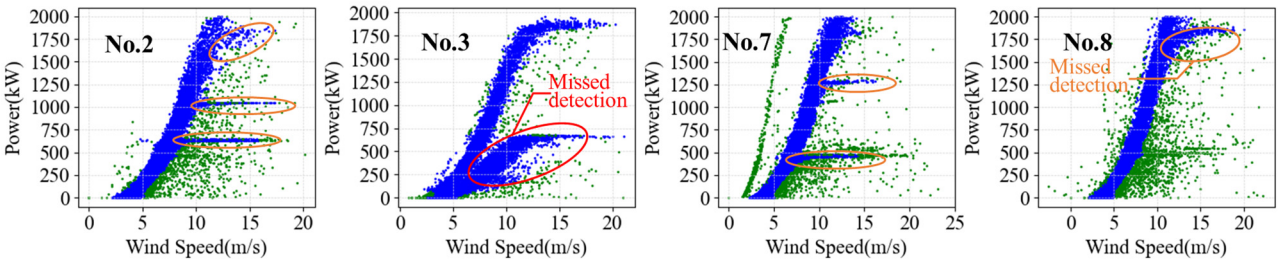

Figure 6. The results of abnormal data cleaning using the quartile method.

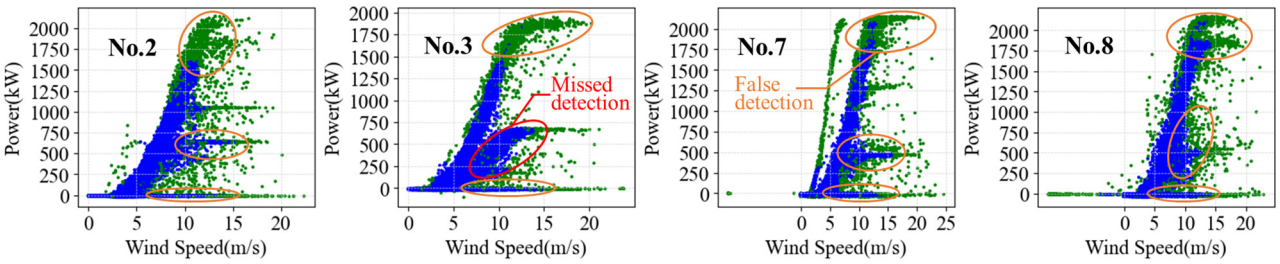

Figure 7. The results of abnormal data cleaning using the Isolation Forest method.

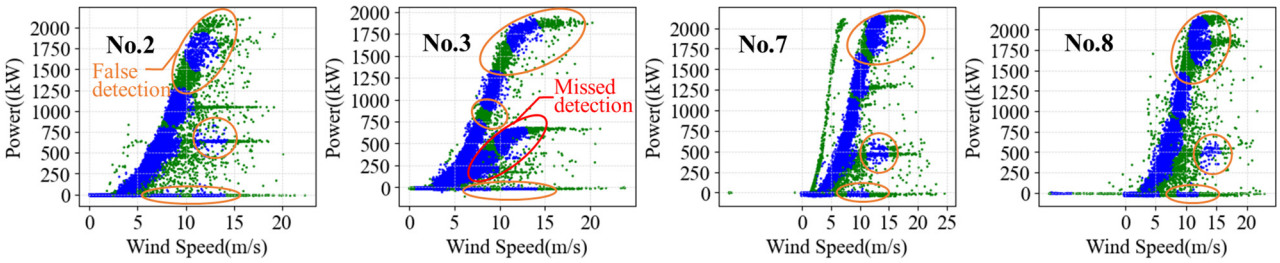

Figure 8. The results of abnormal data cleaning using the k-means method.

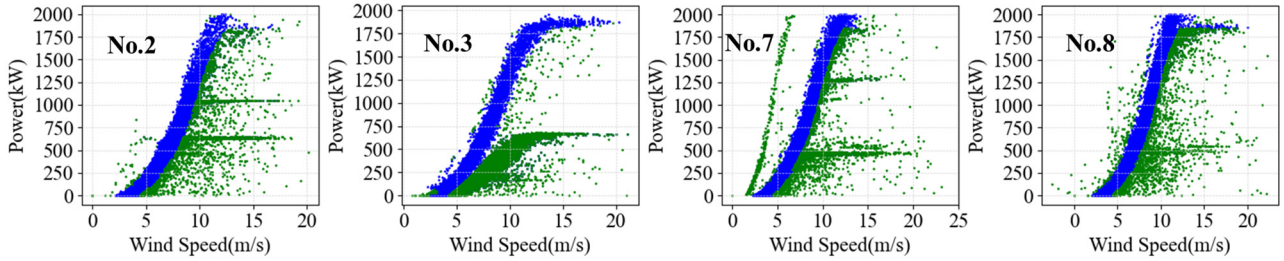

Figure 9. The results of abnormal data cleaning using the quartile RANSAC algorithm.

The interval unit for the quartile algorithm is set to 25 kW, and Figure 6 shows the cleaning results using the quartile method. This process effectively removes a significant amount of abnormal data, but there are still a few scattered abnormal data points remaining on the left side, and there are clusters of abnormal data and a large number of stacked abnormal data on the right side.

For the Isolation Forest algorithm, the contamination ratio is set to 0.05, and the maximum number of features (max_features) is set to 0.1. Isolation Forest requires two prerequisites: a small proportion of abnormal data and large differences in data features. As can be seen in Figure 7, Isolation Forest is not suitable for global detection but is only effective in eliminating sparse, abnormal data and stacked abnormal data that are far from normal values.

The k-means algorithm is set with a cluster center count of 10. As shown in Figure 8, the data cleaning results using the k-means algorithm still exhibit several shortcomings. There are misdetections on the wind power curve, and there are missed detections within the clusters of abnormal data on the right side. The selection of parameters, such as the number of clusters and cluster centers in the k-means algorithm, directly affects the clustering effect.

The data cleaning results of the proposed quartile RANSAC algorithm, as shown in Figure 9, reveal that it achieves relatively satisfactory results for four different types of wind turbines with abnormal data. The high degree of consistency between the cleaned data and the standard power curve demonstrates the effectiveness of the quartile RANSAC algorithm.

4.3.2. Evaluation Metrics

To further quantitatively analyze the performance of abnormal data cleaning algorithms, this paper introduces data deletion rate R (%), cleaning time T (s), mean absolute error (MAE), and root mean square error (RMSE) as evaluation metrics for comparison.

The data deletion rate R (%) is defined as the ratio of abnormal data to the original data, as shown in Equation (6), where n represents the number of abnormal data points and N represents the total number of original data points.

$$R = \frac{n}{N} \times 100\% \tag{6}$$

To more accurately evaluate the model's degree of fit, the cleaned data are segmented into intervals of 0.5 m/s along the standard power curve. The deviation from the standard power curve in each interval is quantitatively expressed by calculating the MAE and RMSE within each interval. The calculation formulas are shown in Equations (7)–(10), where N_i is the amount of data in the ith interval; M is the number of wind speed intervals, P_i is the value of the standard power curve within the ith interval, j represents the jth data point within the ith interval, and Cap is the capacity of the wind turbine generator.

$$MAE_i = \frac{1}{N_i} \sum_{j=1}^{N_i} |P_i - P_{i,j}| \tag{7}$$

$$RMSE_i = \frac{1}{Cap}\sqrt{\frac{1}{N_i}\sum_{i=1}^{N_i}\left(P_i - P_{i,j}\right)^2} \quad (8)$$

$$MAE = \frac{1}{Cap}\sum_{i=1}^{M} MAE_i \quad (9)$$

$$RMSE = \frac{1}{Cap}\sum_{i=1}^{M} RMSE_i \quad (10)$$

Based on these evaluation metrics, the four abnormal data cleaning algorithms are evaluated, and the results are presented in Figure 10, Tables 1 and 2. Moreover, in Figure 10, the best performance index of each unit is marked in red.

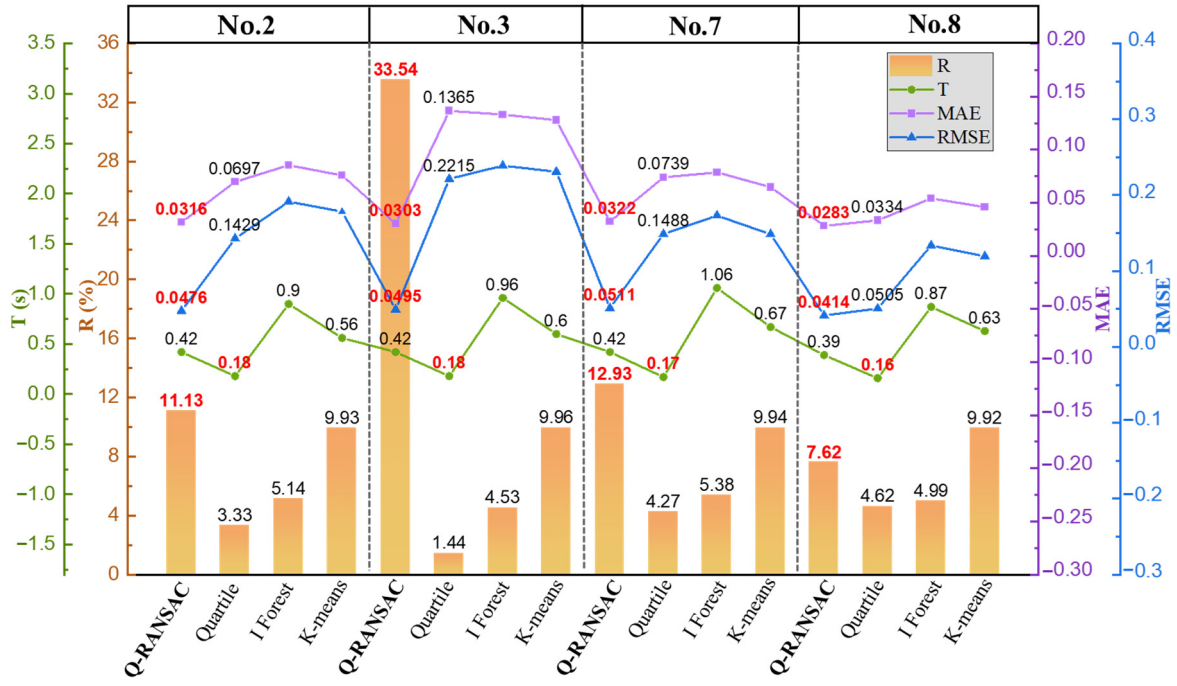

Figure 10. The cleaning results and evaluation metrics of the four algorithms.

Table 1. Comparison of data deletion rate and cleaning time among the four algorithms.

Wind Turbine	Total Number of Data	Quartile RANSAC		Quartile		Isolation Forest		K-Means	
		R (%)	T (s)	R (%)	T (s)	R (%)	T (s)	R (%)	T (s)
No. 2	38,855	11.13	0.42	3.33	0.18	5.14	0.90	9.93	0.56
No. 3	38,995	33.54	0.42	1.44	0.18	4.53	0.96	9.96	0.60
No. 7	43,324	12.93	0.42	4.27	0.17	5.38	1.06	9.94	0.67
No. 8	38,470	7.62	0.39	4.62	0.16	4.99	0.87	9.92	0.63

Table 2. Comparison of MAE and RMSE results among the four algorithms.

Wind Turbine Number	Original Data		Quartile RANSAC		Quartile		Isolation Forest		K-Means	
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
No. 2	0.0982	0.2158	0.0316	0.0476	0.0697	0.1429	0.0853	0.1916	0.0758	0.1783
No. 3	0.1437	0.2582	0.0303	0.0495	0.1365	0.2215	0.1330	0.2389	0.1277	0.2309
No. 7	0.0912	0.1998	0.0322	0.0511	0.0739	0.1488	0.0786	0.1731	0.0648	0.1484
No. 8	0.0671	0.1665	0.0283	0.0414	0.0334	0.0505	0.0542	0.1338	0.0459	0.1193

As shown in Table 1, among the four algorithms, the proposed quartile RANSAC algorithm exhibits a significantly higher data deletion rate compared to the others, with 33.54% and 7.62% for Wind Turbines No. 3 and No. 8, respectively. Combining the wind speed-power scatter plots, Wind Turbine No. 3 has the most abnormal data, while No. 8 has the least, which is consistent with the data deletion rates of the proposed algorithm, proving that it can filter out more abnormal data, regardless of whether they are stacked or sparse. In terms of cleaning time T (s), the quartile method has the shortest cleaning time but a relatively lower data deletion rate, indicating weaker anomaly detection capability. Although the proposed method's cleaning time is slightly longer than the quartile method, it outperforms Isolation Forest and k-means. The proposed method ensures cleaning efficiency while significantly optimizing the cleaning effect and making the data deletion rate more reasonable.

As seen in Table 2, the proposed quartile RANSAC algorithm achieves the smallest MAE and RMSE values. For wind turbine No. 3, which has a high proportion of stacked anomalies, the MAE and RMSE of the proposed algorithm are 0.0303 and 0.0495, respectively, representing a 78% reduction in both MAE and RMSE compared to the quartile method, which performed the best among the compared algorithms. Compared with the original data, MAE and RMSE are reduced by 79% and 81%, respectively. This demonstrates the significant superiority of the quartile RANSAC algorithm in cleaning data with a high proportion of stacked anomalies. For wind turbines No. 2, 7, and 8 with other types of abnormal data, the proposed method reduces MAE by 54%, 56%, and 15%, respectively, and RMSE by 67%, 66%, and 18%, respectively, compared to the quartile method. This further proves the outstanding cleaning performance of the quartile RANSAC algorithm in handling both high-proportion stacked anomalies and various other types of anomalies, verifying its superiority and wide applicability.

The abnormal data cleaning method for wind turbines proposed in this paper provides strong support for wind power prediction, state monitoring, and fault diagnosis in wind power generation. First, the cleaned data are more accurate and reliable, providing high-quality data support for wind power prediction. Secondly, it can help with state monitoring. By analyzing the cleaned data, the operating state of wind turbines can be understood more clearly. Thirdly, it lays the foundation for fault diagnosis. When there are abnormal fluctuations in the data, the key components of the fan can be checked specifically to see if there are fault risks.

5. Conclusions

(1) A novel method for abnormal data cleaning based on a classification processing framework is proposed, which employs operational guidelines, the quartile method, and the RANSAC regression algorithm for three types of abnormal data. This staged approach significantly enhances the robustness and accuracy of cleaning data with a high proportion of stacked anomalies.

(2) Through a case study on the cleaning of abnormal data from wind turbines with a high proportion of stacked abnormalities, the high degree of consistency between the cleaned data and the standard power curve indicates that the cleaning effect is good. Furthermore, the proposed method is accurate for cleaning other types of abnormal wind turbines, and the effectiveness of the quartile RANSAC algorithm has also been proved.

(3) To validate the significant advantages of the proposed method, it was compared with quartile, isolation forest, and k-means algorithms. The cleaning results intuitively demonstrate that the proposed method significantly outperforms the other three existing algorithms in terms of cleaning effectiveness. The introduction of evaluation metrics further accurately demonstrates the superiority of the cleaning results, with the proposed method achieving a more reasonable data deletion rate and excellent cleaning efficiency. Compared to the quartile method, which performed the best among the compared algorithms, the proposed method reduces MAE by 54%, 78%,

56%, and 15% and RMSE by 67%, 78%, 66%, and 18%, respectively, across the four wind turbines, proving the better performance of the quartile RANSAC algorithm in abnormal data cleaning. It must be pointed out that the performance of the algorithm proposed in this paper depends to some extent on its parameter configuration. If the parameters are not set reasonably, it may lead to deviations in the results. The authors will strive to address this issue in future work.

Author Contributions: Conceptualization, X.Z. and F.Z.; methodology, F.Z.; software, F.Z. and Y.L.; validation, F.Z. and Z.L.; formal analysis, Z.X. and Z.L.; investigation, K.D. and Y.L.; resources, X.Z.; data curation, Z.X. and F.Z.; writing—original draft preparation, F.Z.; writing—review and editing, X.Z., Z.X. and K.D.; visualization, K.D. and Y.L.; supervision, X.Z.; project administration, X.Z.; funding acquisition, X.Z. and Z.X. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by "Natural Science Foundation of Henan Province" grant number "202300410117", "Natural Science Foundation of Henan Province" grant number "242102240129", and "China Postdoctoral Science Foundation" grant number "2022M712382".

Data Availability Statement: The data that support the findings of this study are available from the corresponding author upon reasonable request.

Acknowledgments: We express our gratitude to the funding project mentioned in "Funding" for its sponsorship of this research. Additionally, we also extend our thanks to the following projects for their support: the Science Foundation of Henan University of Technology (Grant No. 31401252).

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

References

1. Global Wind Energy Council, Global Wind Report 2024 (Global Wind Energy Council, 2024). Available online: https://gwec.net/global-wind-report-2024/ (accessed on 1 July 2024).
2. Li, Z.; Jiang, Y.; Guo, Q.; Hu, C.; Peng, Z. Multi-dimensional variational mode decomposition for bearing-crack detection in wind turbines with large driving-speed variations. *Renew. Energy* **2018**, *116*, 55–73. [CrossRef]
3. Xu, Y.; Jia, L.; Yang, W. Correlation based neuro-fuzzy Wiener type wind power forecasting model by using special separate signals. *Energy Convers. Manag.* **2022**, *253*, 115173. [CrossRef]
4. Tian, B.; Zhang, Y. Energy storage operation control strategy for smoothing wind power based on multi-objective cooperative game. *High Voltage Eng.* **2023**, *49*, 2546–2557.
5. Nielson, J.; Bhaganagar, K.; Meka, R.; Alaeddini, A. Using atmospheric inputs for Artificial Neural Networks to improve wind turbine power prediction. *Energy* **2020**, *190*, 116273. [CrossRef]
6. McKinnon, C.; Carroll, J.; McDonald, A.; Koukoura, S.; Plumley, C. Investigation of isolation forest for wind turbine pitch system condition monitoring using SCADA data. *Energies* **2021**, *14*, 6601. [CrossRef]
7. Maldonado-Correa, J.; Martín-Martínez, S.; Artigao, E.; Gómez-Lázaro, E. Using SCADA data for wind turbine condition monitoring: A systematic literature review. *Energies* **2020**, *13*, 3132. [CrossRef]
8. Yang, W.; Tavner, P.J.; Crabtree, C.J.; Feng, Y.; Qiu, Y. Wind turbine condition monitoring: Technical and commercial challenges. *Wind Energy* **2014**, *17*, 673–693. [CrossRef]
9. Liu, J.; An, B.; Zhang, W.; Gan, Q. Review of health condition evaluation of large wind turbines. *Power Syst. Prot. Control* **2023**, *51*, 176–187.
10. Wen, X.; Xu, Z. Wind turbine fault diagnosis based on ReliefF-PCA and DNN. *Expert Syst. Appl.* **2021**, *178*, 115016. [CrossRef]
11. Elusakin, T.; Shafiee, M. Fault diagnosis of offshore wind turbine gearboxes using a dynamic Bayesian network. *Int. J. Sustain. Energy* **2022**, *41*, 1849–1867. [CrossRef]
12. Lou, J.; Xu, J.; Lu, H.; Qu, C.; Li, S.; Liu, R. Wind turbine data cleaning algorithm based on power curve. *Autom. Electr. Power Syst.* **2016**, *40*, 116–121.
13. Zheng, L.; Hu, W.; Min, Y. Raw wind data preprocessing: A data-mining approach. *IEEE Trans. Sustain. Energy* **2014**, *6*, 11–19. [CrossRef]
14. 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.
15. Hou, G.; Wang, J.; Fan, Y. Wind power forecasting method of large-scale wind turbine clusters based on DBSCAN clustering and an enhanced hunter-prey optimization algorithm. *Energy Convers. Manag.* **2024**, *307*, 118341. [CrossRef]
16. 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. [CrossRef]

17. 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]
18. Ye, X.; Lu, Z.; Qiao, Y.; Min, Y.; O'Malley, M. Identification and Correction of Outliers in Wind Farm Time Series Power Data. *IEEE Trans. Power Syst.* **2016**, *31*, 4197–4205. [CrossRef]
19. 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]
20. 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.
21. 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]
22. Li, L.; Liang, Y.; Lin, N.; Yan, J.; Meng, H.; Liu, Y. Wind speed cleaning method for wind turbine considering spatial-temporal correlation. *Acta Energy Sol. Sin.* **2024**, *45*, 461–469.
23. 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]
24. Wang, Z.; Wang, L.; Huang, C. A fast abnormal data cleaning algorithm for performance evaluation of wind turbine. *IEEE Trans. Instrum. Meas.* **2020**, *70*, 5006512. [CrossRef]
25. 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]
26. Chen, H.; Chen, J.; Han, G.; Cui, Q. Winding down the wind power curtailment in China: What made the difference? *Renew. Sustain. Energy Rev.* **2022**, *167*, 112725. [CrossRef]
27. Fischler, M.A.; Bolles, R.C. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM* **1981**, *24*, 381–395. [CrossRef]