energies

MDPI

Article

SCADA Data-Based Working Condition Classification for Condition Assessment of Wind Turbine Main Transmission System

Huanguo Chen [1,*], Chao Xie [1], Juchuan Dai [2], Enjie Cen [3] and Jianmin Li [1]

[1] Faculty of Mechanical Engineering and Automation, Zhejiang Sci-Tech University, Hangzhou 310018, China; xc_062966@163.com (C.X.); ljmzrz@126.com (J.L.)
[2] School of Mechanical Engineering, Hunan University of Science and Technology, Xiangtan 411201, China; daijuchuan@163.com
[3] Hangzhou Dingchuan Information Technology Co., Ltd., Hangzhou 310016, China; cenenjie@126.com
* Correspondence: chen8025@126.com

Abstract: Due to the complex and variable conditions under which wind turbines operate, existing working condition classification methods are inadequate for condition assessment of the main transmission system. Because working conditions are too few after classification, it cannot effectively describe the complex and variable working conditions of wind turbine. This can lead to high false-alarm rates in the condition monitoring, which affect normal operations. This paper proposes a working condition classification method for the main transmission system of wind turbines based on supervisory control and data acquisition (SCADA) data. Firstly, correlation analysis of SCADA data acquired by wind farm is used to select the parameters relevant to the main transmission system. Secondly, according to the wind turbine control principle, the working conditions are initially divided into four phases: shutdown, start-up, maximum wind energy tracking, and constant speed. The *k*-means clustering algorithm is used to subdivide the maximum wind energy-tracking phase and constant speed phase, which account for a larger proportion of the working conditions, to achieve better classification. Finally, a case study is used to demonstrate the calculation of alarm thresholds and alarm rates for each working condition. The results are compared with the direct use of *k*-means clustering for working condition classification. It is concluded that the proposed method can significantly reduce the false-alarm rate of the vibration detection process.

Keywords: main transmission system; working conditions classification; wind turbine working characteristics; alarm threshold; *k*-means clustering; SCADA data

Citation: Chen, H.; Xie, C.; Dai, J.; Cen, E.; Li, J. SCADA Data-Based Working Condition Classification for Condition Assessment of Wind Turbine Main Transmission System. *Energies* **2021**, *14*, 7043. https://doi.org/10.3390/en14217043

Academic Editors: Davide Astolfi and Francesco Castellani

Received: 8 September 2021
Accepted: 22 October 2021
Published: 28 October 2021

1. Introduction

Double-fed wind turbine main transmission systems are mainly composed of a hub, main shaft, bearings, gearbox, and couplings. They not only provide a transmission channel for energy but also a transmission path for the unit load. Under long-term operation in harsh working environments, such as extreme climatic conditions, which involve a wide range of working speeds and a wide excitation band, time-varying wind loads often cause oscillation in wind turbine towers, which are somewhat flexible. Then, the inertial forces of the components of the transmission system may combine with the system's pneumatic load and increase the degree of force fluctuations in the components, leading to increased failure rates. Failures of the main transmission system account for 40–60% of wind turbine downtime [1]. One of the causes of abnormal transmission system conditions is that the real-time dynamic loads and the dynamic behaviour of the transmission system under diverse working conditions remain unclear. The operating environments of the main transmission system are usually complex and harsh. Different wind speeds, wind directions, and rotor speeds have different effects on the energy output of a wind turbine [2], which makes it

very difficult to monitor the condition of its main transmission system. Therefore, it is essential to classify the working conditions reasonably and accurately before conducting a conditions assessment.

At present, there are few studies on classifying the working conditions of wind turbine main transmission systems. Most consider the working conditions of the whole wind turbine and classify the working conditions according to the operating characteristics of the wind turbine and the use of clustering algorithms. Yan [3] and Mei [4] divided the wind turbine speed and torque control process into four phases: start-up, maximum wind energy tracking, constant speed generation, and constant power generation. Dai et al. [5] classified wind turbine operation according to the relationship between rotor speed and power as start-up, first transition, maximum power point tracking, second transition, and full power. Yang et al. [6] classified wind turbine operation according to power curve parameters into four phases. Cheng et al. [7] classified wind turbine operating conditions based on wind speed-power curves, rotational speed-power curves, rotational speed-pitch angle curves, and power-pitch angle curves into seven states: normal operation, normal shutdown, power-limited operation, adjustment process, off-grid process, abnormal shutdown, and on-grid acceleration. Ling [8] identified the working conditions of the main bearing according to wind speed and classified the working conditions as low power, rapid power increase, and stable power. Gu et al. [9] used statistical analysis to classify the working conditions according to ambient temperature first and then by wind speed. However, the number of classes was relatively small and does not effectively describe the complex and variable working conditions of wind turbines. This causes a high false-alarm rate during the monitoring of actual conditions.

Alternatively, working conditions have been classified using clustering algorithms. Dong et al. [10], Xing et al. [11], Zhang et al. [12], and Wang et al. [13] selected the characteristic parameters of working conditions by analysing the correlations in SCADA data and using the *k*-means clustering algorithm to classify historical working conditions during normal wind turbine operation. Liu et al. [14] proposed semi-supervised *k*-means clustering with stream-wise distance as a similarity measure for analysing massive SCADA datasets to classify wind turbine working conditions. Jin [15] and Liu et al. [16] selected working conditions feature parameters and used the fuzzy C-means (FCM) clustering algorithm to classify the operating conditions. Yin et al. [17] and Chen et al. [18] used a Gaussian mixture model (GMM)-based clustering method to classify wind turbine operating conditions based on condition monitoring data. Ma et al. [19] proposed the adaptive classification of conditions using GMM. Wang et al. [20] selected three parameters—wind speed, generator speed, and active power—and applied the GMM clustering algorithm to classify training data into three sub-conditions. Han. [21] divided SCADA data by month and determined the optimal interval length according to particle swarm optimization (PSO) and then carried out interval clustering. Han et al. [22] proposed a clustering method for wind farms based on correlation analysis and significance testing (CA-ST). Zheng et al. [23] proposed a PSO optimization kernel principal element analysis (KPCA) method for classifying the working conditions of offshore wind turbines. However, most of these studies are based on the direct application of a clustering algorithm to data and do not consider the wind turbines' operating mechanisms. Hence, the correlations between wind turbine operating parameters and operating conditions are unclear and lack theoretical support.

In the field of machine learning, clustering is a classical unsupervised learning method. Commonly used clustering algorithms and their advantages and disadvantages are summarised in Table 1.

Based on a comparative analysis of the clustering algorithms in Table 1, the *k*-means clustering algorithm was selected to cluster the working conditions of the main transmission system of a wind turbine due to the large number of dimensions in the large amount of SCADA data. The Calinski–Harabasz (CH) criterion was used to determine the number of clusters in the *k*-means clustering algorithm, while the initial cluster centres were selected by the *k*-means++ algorithm. The specific method is described in Section 3.1.2.

The working conditions of a wind turbine's main transmission system are relatively complex. Figure 1 shows the variation in the vibrational acceleration of the main shaft of a 2.5-MW wind turbine according to wind speed and power. As can be seen, even when an operating parameter is fixed, the vibration still shows a very high variability [24]. In classifying the main transmission system's working conditions, there is a small number of conditions after classification, and the vibrational acceleration may change drastically within each condition. Therefore, it is not possible to determine whether the vibrational acceleration changed drastically due to a fault in the main transmission system fault, and it is difficult to make an accurate judgment of the system's operating status.

Table 1. Commonly used clustering algorithms and their advantages and disadvantages.

Category	Representative Algorithms	Advantages	Disadvantages
Partition-based clustering	*k*-means [25]	Simple and fast. Scalable and efficient for handling large data sets.	*k*-values are difficult to estimate. The choice of initial centroids can affect the clustering results to a large extent.
Hierarchy-based clustering	Balanced Iterative Reducing and Clustering Using Hierarchies (BIRCH) [26]	No need to enter the number of categories K. Memory saving and fast clustering. Noise points can be identified, and the dataset can be pre-processed for initial classification.	Not suitable for clustering data with high-dimensional features. Complex adjustment of key parameters has a large impact on the final result.
Density-based clustering	Density-Based Spatial Clustering of Applications with Noise (DBSCAN) [27]	No need to determine *k*-values in advance. Arbitrarily shaped clusters can be found. Outliers can be identified. Initial centroids do not affect clustering results.	Not suitable for clustering high-dimensional data. Not suitable for clustering data with changing density. Difficult to determine optimal values for parameters.
Network-based clustering	Statistical Information Grid (STING) [28]	Fast clustering	Parameter-sensitive, unable to handle irregularly distributed data. Low accuracy of clustering results.
Model-based clustering	Self-Organized Maps (SOM) [29], (Gaussian Mixture Model) GMM [30]	The classification of "classes" is expressed in probabilistic form and the characteristics of each class can be expressed in terms of parameters.	Inefficient execution, especially when the number of distributions is large, and the amount of data is small.
Fuzzy-based clustering	Fuzzy c-means (FCM) [31]	Classification according to the principle of maximum subordination in fuzzy sets. Better for clustering normally distributed data.	Dependent on initial clustering centres. Longer clustering time for larger data volumes. No guarantee of convergence to an optimal solution.
Graph-based clustering	Spectral clustering [32]	Spectral clustering works well when there are few clustering categories. Suitable for high-dimensional clustering. Has the ability to cluster on an arbitrarily shaped sample space and converge to a globally optimal solution.	Very sensitive to the choice of clustering parameters. Only applicable to balanced classification problems.

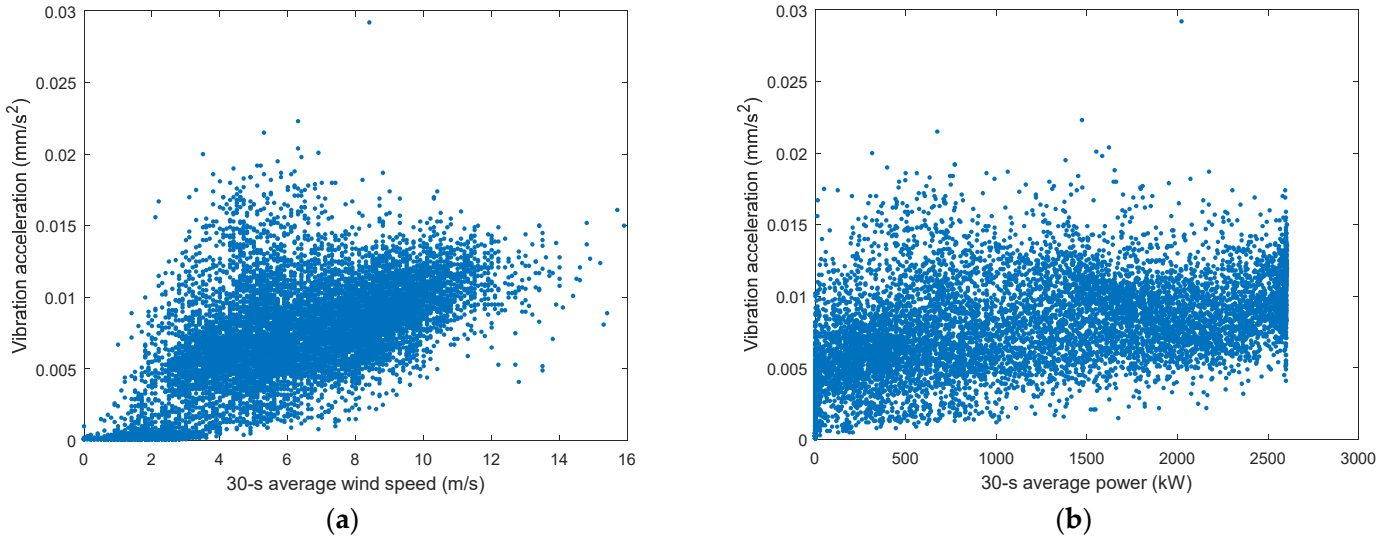


Figure 1. Variations in spindle vibrational acceleration with (**a**) 30-s average wind speed and (**b**) 30-s average power.

To address the above problems, this paper proposes a method for classifying a wind turbine main transmission system's conditions based on analysis of the wind turbine's operating characteristics and use of a *k*-means clustering algorithm. The main contributions of this paper are as follows:

- A transmission system working condition classification based on wind turbine operating characteristics and a *k*-means clustering algorithm is proposed. It can solve the problems of traditional classification systems, such as classes being insufficient in clarity or number and high false-alarm rates in the main drivetrain vibration detection process.
- A method for selecting the status parameters of the main transmission system based on correlation analysis is proposed. It avoids the influence of feature parameter omissions in the process of selecting feature parameters and improves the validity of SCADA data.
- During vibration monitoring, the false-alarm rate is used as an index to verify the validity of the transmission system's working condition classification.

The remaining parts of the paper are organized as follows. The main transmission system's status feature parameters are selected according to correlation analysis in Section 2. Classification of the transmission system's historical working conditions based on turbine operating characteristics is combined with a clustering algorithm in Section 3. The false-alarm rate in vibration detection is used as an index for the evaluation of classification results in Section 4. A case study is used to validate the proposed method in Section 5. Conclusions are drawn in Section 6. A flowchart of the working condition classification system is shown in Figure 2.

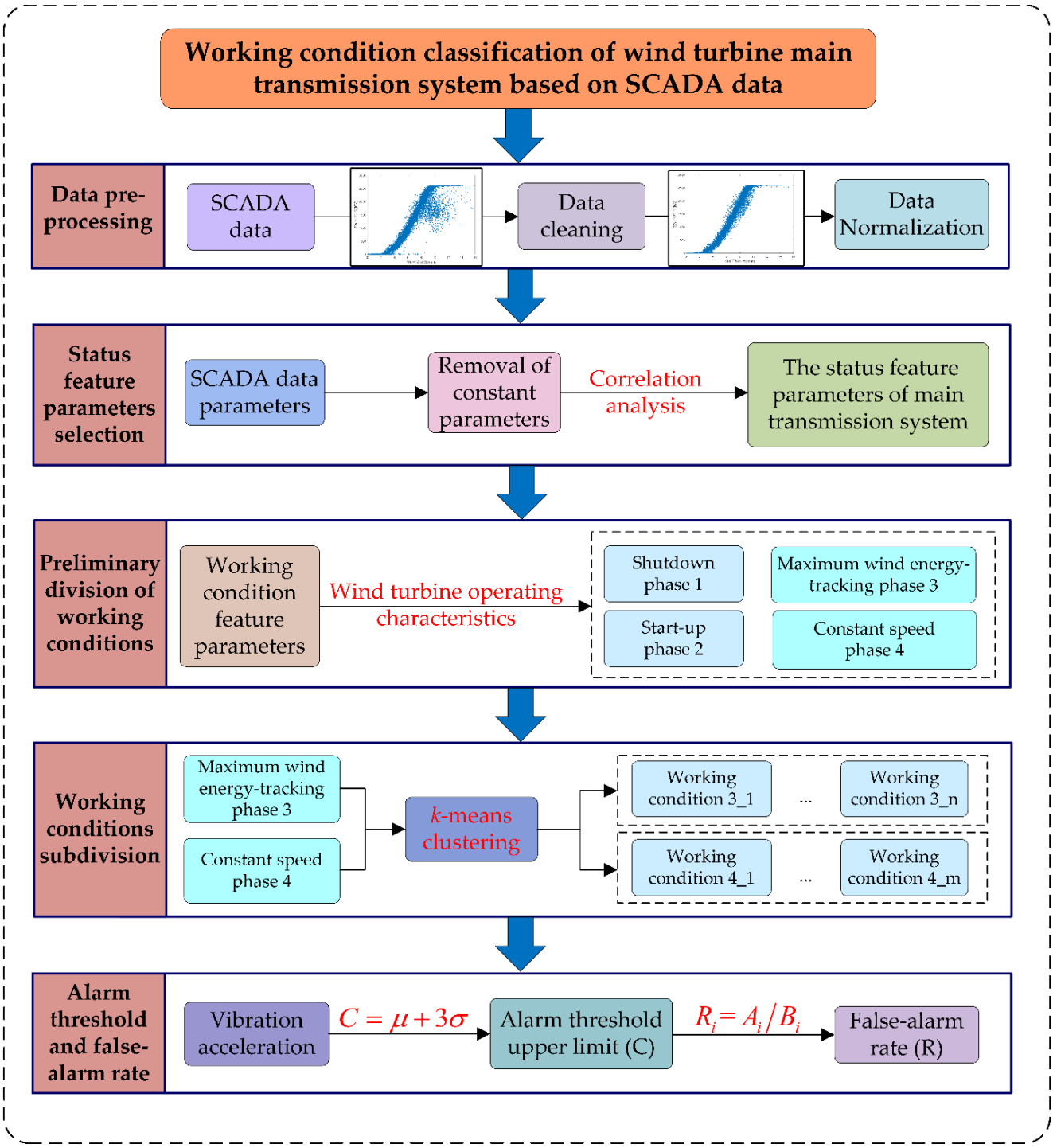


Figure 2. Flowchart of the working condition classification system.

2. Selection of the Main Transmission System's Status Feature Parameters

The main transmission system of wind turbine includes a hub, main shaft, bearings, gearbox, and couplings. The SCADA system collects a large amount of data with a large number of dimensions. The data contains many redundant variables, so in condition assessment, it is necessary to eliminate variables that are unrelated to the main transmission system. The Pearson correlation coefficient is used to select the characteristic parameters of the main transmission system's operating status from the original SCADA data.

The Pearson correlation coefficient is used to measure the degree of linear correlation between two random variables. Assuming the existence of variables X and Y, the Pearson correlation coefficient between them is:

$$\rho_{X,Y} = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E((X-\mu_X)(Y-\mu_Y))}{\sigma_X \sigma_Y} \quad (1)$$

where $\mathrm{cov}(X,Y)$ is the covariance of X and Y, and σ_X, σ_Y are their standard deviations.

If the total number of X and Y samples is n, then the correlation coefficient between X and Y is:

$$r = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n} (X_i - \overline{X})^2} \sqrt{\sum_{i=1}^{n} (Y_i - \overline{Y})^2}} \quad (2)$$

The Pearson correlation coefficient r quantifies the linear correlation between the variables X and Y with the range [–1, 1]. Higher values indicate a stronger correlation. In general, it is considered that $|r| \geq 0.5$ is necessary to consider that two variables are correlated [33].

The SCADA system collects a large amount of data with a large number of dimensions. This includes some constant parameters, such as power factor, pitch standby supply voltage, and grid-side voltage, which do not express the operating status of the wind turbine at the site and therefore need to be eliminated. Based on expert experience [33], a total of 18 status parameters related to the operating state of the main transmission system were selected from the SCADA system, as shown in Table 2. Here, the average 30-s power is used as the performance index. Status parameters with a strong relationship to the active power are selected from the SCADA system and the Pearson correlation coefficient is calculated, as shown in Table 2.

Table 2. Pearson correlation coefficients between transmission system status parameters and the average 30-s power.

Status Parameter	Pearson Correlation Coefficient	Status Parameter	Pearson Correlation Coefficient
Average spindle speed	0.8072	Average wind speed	0.9256
Average torque	0.9953	30-s average wind speed	0.9263
Average wind direction	0.0396	600-s average wind speed	0.9171
Average outdoor temperature	−0.1739	600-s average power	0.9822
Average 30-s wind direction	0.0403	Average cabin temperature	−0.2198
Average gearbox oil temperature	0.4117	Average cabin cabinet temperature	−0.1544
Average gearbox high-speed end-bearing temperature	0.7873	Average main bearing temperature	0.6464
Average gearbox oil distributor outlet pressure	0.7491	Average gearbox low-speed end bearing temperature	0.6206
Average gearbox oil filter inlet pressure	0.8267	Average spindle vibration acceleration	0.8133

From the correlation analysis, the state parameters in Table 2 with correlation coefficients $|r| \geq 0.5$ were selected as the status parameters of the main transmission system. Among the average wind speed, 30-s average wind speed, and 600-s average wind speed, the 30-s average wind speed had the highest correlation with the active power and so was selected as the wind speed status parameter. The 30-s average power was selected as the power parameter rather than the 600-s average power. The average main shaft speed, 30-s average wind speed, 30-s average power, average torque, average main bearing temperature, average gearbox high-speed end-bearing temperature, average gearbox low-speed end-bearing temperature, average main shaft vibration acceleration, average gearbox oil filter inlet pressure, and average gearbox oil distributor outlet pressure were selected as the 10 status parameters of the main transmission system.

3. Classification of the Working Conditions of the Main Transmission System

3.1. Introduction to the Principles of the k-Means Clustering Algorithm

The *k*-means clustering algorithm is a clustering method based on the similarity between samples. The objective is to minimize the sum of squares of the distances between all samples and their corresponding cluster centres.

Assuming an input sample set $X = \{x_1, x_2, \cdots, x_n\}$, the number of clusters is k. *k*-means clustering can be expressed as the sum of squared errors (*SSE*) of convergence.

$$SSE = \sum_{j=1}^{k} \sum_{\mu_i \in \varphi_j} \|\mu_i - \mu_j\|^2 \tag{3}$$

where φ_j is the first j cluster; μ_i is the sample point in φ_j; and μ_j is φ_j the centre of mass.

3.1.1. The *k*-Means Clustering Algorithm Process

The *k*-means clustering algorithm process is as follows:

1. Generate k initial centre-of-mass points using the *k*-means++ algorithm: $\{\mu_1, \mu_2, \cdots, \mu_k\}$.
2. Calculate the distance between each sample point and the centre-of-mass points.
3. Assign sample points to the class nearest to them.
4. Calculate the centre of mass of each class using the sample points that have just been grouped: calculate the mean value of each cluster coordinate as the centre of mass.
5. Repeat steps 3–5 until its centre of mass no longer changes, or the maximum number of iteration steps is reached.
6. Output the cluster division $C = \{C_1, C_2, \cdots, C_k\}$.

3.1.2. The *k*-Means++ Algorithm Process

The *k*-means++ algorithm for generating the initial cluster of centre-of-mass points flows is as follows:

1. Create an empty set S for storing the k prime points of the cluster.
2. Select a random instance from the sample set X called μ_1, and add it to the first cluster S as the centre of mass.
3. For each instance x_i in the dataset X, calculate the square of the distance to the centre of mass of each cluster within dataset S, the smallest of which is the square of the distance x_i to S:
$$d(x_i, S)^2 = \min_{\mu_j \in S} \|x_i - \mu_j\|^2 \tag{4}$$
4. The probability of each sample being selected as the next cluster centre is calculated as follows. The next cluster centre point μ_i is selected by the roulette wheel method and added to S.
$$P(x_i) = \frac{d(x_i, S)^2}{\sum_j d(x_j, S)^2} \tag{5}$$
5. Repeat steps 3–4 until k clusters of centre-of-mass points have been selected.

3.1.3. Determination of the Number of Clusters k

In cluster analysis, the number of clusters k needs to be obtained first, which determines the effectiveness of clustering. If the k value is too small, the range of each condition interval is too large. This can easily lead to the classification of the optimal working interval range being too large or there being no working conditions that satisfy the minimum support and minimum feasible degree in the classification results, resulting in failure of classification. If k is too large, the interval of each parameter is too detailed, which increases the dimensionality of the parameters and increases the intermediate data of the mining process, resulting in a reduction in mining speed. Therefore, this paper adopts the Calinski–Harabasz (CH) criterion to determine a reasonable number of clusters k.

The CH criterion, also known as the variance ratio criterion, assesses the effect of clustering using the degree of denseness within clusters and the degree of dispersion between clusters, which is calculated as follows:

$$S(k) = \frac{tr(B_k)}{tr(W_k)} \times \frac{m-k}{k-1} \tag{6}$$

where m is the total number of samples in the training set, k is the number of clusters, B_k is the covariance matrix between categories, W_k is the covariance matrix of the data within categories, and tr is the trace of the matrix.

After the clustering is completed, the smaller the covariance of the data within categories and the larger the covariance between categories (i.e., the tighter the classes are themselves and the more dispersed the classes are), the better the clustering effect will be, and, accordingly, the higher the CH score will be. Therefore, the optimal number of clusters can be determined by calculating the CH scores at different k values.

3.2. Main Transmission System Working Conditions Classification Based on the k-Means Clustering Algorithm

Most traditional working conditions are classified directly by clustering algorithms after selecting the working parameters, without taking into account the operating characteristics of the wind turbine. In different wind speed ranges and under different working conditions, wind turbines are adjusted by different control methods, so it is necessary to analyse their control strategies and classify the working conditions in different stages.

Current mainstream doubly-fed wind turbines have a variable-speed and constant-frequency control strategy; i.e., the wind turbine speed varies with the wind speed to maintain the best blade-tip-speed ratio and maximum wind-energy conversion efficiency. Figure 3 shows the relationships between the power of a variable-speed, constant-frequency wind turbine, and wind speed and wind turbine speed. According to the wind turbine's power characteristics and wind speed variation, the operation can be theoretically divided into five phases as follows [34]:

1. Shutdown phase (OA and E+): Wind speeds are less than the cut-in wind speed v_{cut_in} or greater than the cut-out wind speed v_{cut_out};
2. Start-up phase (AB): Wind speeds are greater than the cut-in wind speed v_{cut_in} and less than the wind speed v_1. The wind turbine speed is limited to the minimum speed $\omega_{\min}$;
3. Maximum wind-energy tracking phase (BC): Wind speed is between v_1 and v_2, the wind turbine speed is between the minimum speed $\omega_{\min}$ and the rated speed ω_{rate}, the wind turbine tip speed ratio remains optimal, and the wind energy utilization coefficient remains at the maximum;
4. Constant speed phase (CD): Wind speeds are between v_2 and the rated wind speed v_{rate}, and the wind turbine speed remains at the rated speed ω_{rate};
5. Constant power phase (DE): Wind speeds are between the rated wind speed v_{rate} and cut-out wind speed v_{cut_out}, the wind turbine speed is at the rated speed ω_{rate}, and the wind power utilization coefficient is adjusted by adjusting the pitch angle β so that the wind power output is kept at the rated power P_{rate}.

From the main transmission system status parameters selected by correlation analysis, operating parameters are selected to classify the main transmission system's operating conditions. According to Figure 3, it can be seen that the stages of wind turbine operation are closely related to wind speed, wind turbine speed, and power. The control strategy for each stage is different. Firstly, according to the relationships between power and wind speed and wind turbine speed, there are five categories because the constant speed stage includes the constant power stage. Then, the two stages are combined into one category and collectively referred to as the constant speed stage, making a total of four major categories.

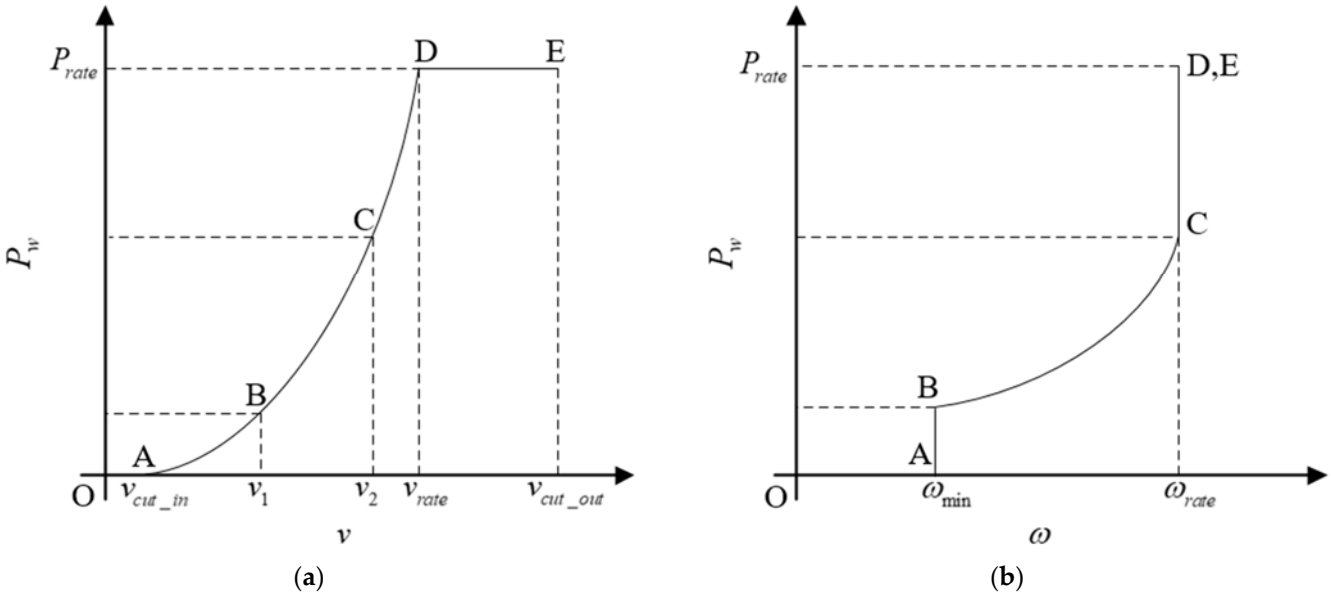


Figure 3. Wind turbine power (P_w) vs. (**a**) wind speed (v) and (**b**) wind turbine speed (ω).

The purpose of the condition classification is to improve the accuracy of later condition assessment by subdividing the operating conditions of the wind turbine. The maximum wind-energy tracking stage and constant speed stage are the normal working stages of the wind turbine and produce a large amount of data, so it is necessary to subdivide the working conditions according to their operating characteristics. In the maximum wind energy tracking stage, the power varies with wind speed and wind turbine speed, so the *k*-means clustering algorithm is used to subdivide the working conditions at this stage according to wind turbine speed, wind speed, and power. In the constant speed stage, the wind turbine speed is constant, so the *k*-means clustering algorithm is used to subdivide the working conditions according to wind speed and power. Finally, we can obtain the final working condition classification results.

4. Determination of Alarm Thresholds

In the process of wind turbine SCADA system detection, the vibrational acceleration of the main transmission system can be detected by sensors to reflect the operating status. A vibrational acceleration alarm is mostly an over-limit alarm; i.e., as long as the threshold is exceeded, the alarm will be generated; otherwise, it will not [35]. The data used in this paper refer to normal wind turbine operation without failure. Theoretically, the detection system should not issue an alarm; however, during actual state monitoring, false alarms are difficult to avoid, which requires adjustment of the alarm threshold to reduce the false-alarm rate as much as possible.

Figure 4 shows the frequency distribution of the average spindle vibrational acceleration from SCADA data. The frequency distribution of Figure 4a contains all vibration data of the wind turbine. In addition, when the wind turbine is running, the vibration acceleration frequency distribution is shown as Figure 4b. Comparing the two graphs, we can see that the peak on the left side of Figure 4a, which appears in the shutdown phase of the wind turbine, we do not consider.

We choose its arithmetic mean μ and standard deviation σ as the basis for the threshold and set the alarm threshold C according to the 3σ criterion.

$$C = \mu + 3\sigma \tag{7}$$

If the vibrational acceleration value exceeds the threshold C, an alarm is generated. To evaluate the reasonableness of this classification, a definition of the false-alarm rate is introduced.

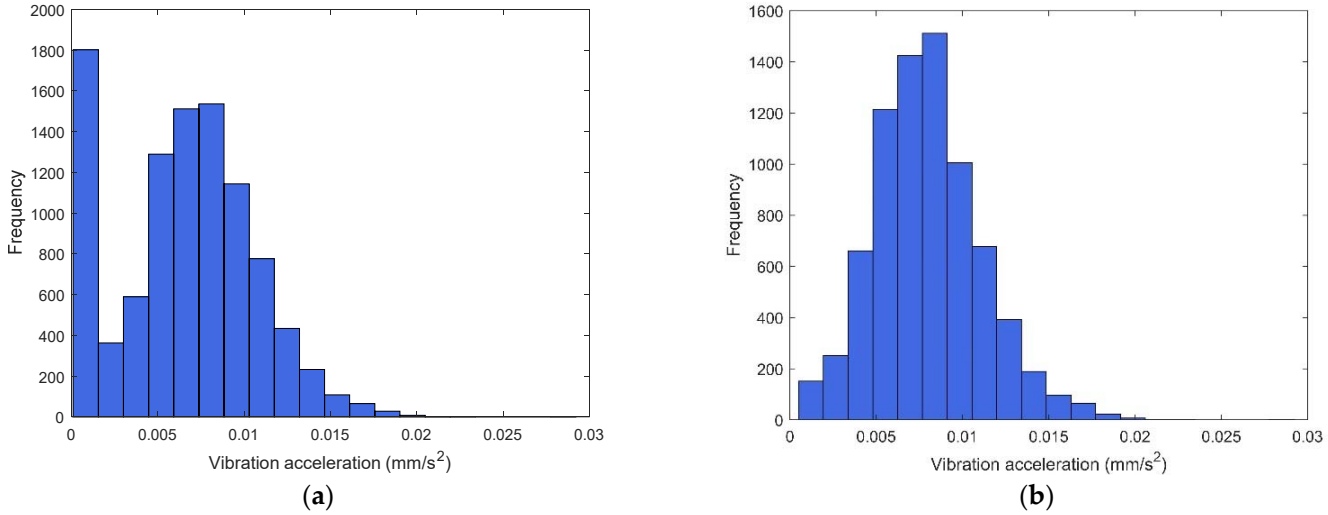


Figure 4. Frequency distribution of vibrational acceleration. (**a**) All vibration data of the wind turbine; (**b**) the vibration data during the wind turbine operation.

In the SCADA historical dataset, it is assumed that the main transmission system has n different working conditions corresponding to n alarm thresholds relating to the characteristic value of vibrational acceleration. For the i-th normal operating conditions, the total number of samples is B_i, and the number of samples exceeding the alarm threshold is A_i. Then, the false-alarm rate R_i for the i-th working conditions is:

$$R_i = A_i / B_i \tag{8}$$

Therefore, the false-alarm rate is used as an index of the merit of the classification system. The lower the rate, the better the classification.

5. A Case Study

5.1. Wind Turbine Overview and SCADA Monitoring Parameters

As a case study, we considered a double-fed wind turbine with a single capacity of 2.5 MW located in a wind farm in northern China. The turbine has been operating normally since its commissioning, with a cut-in wind speed of 3 m/s, a cut-out wind speed of 25 m/s, and a rated wind speed of 9.5 m/s. Its theoretical power curve is shown in Figure 5.

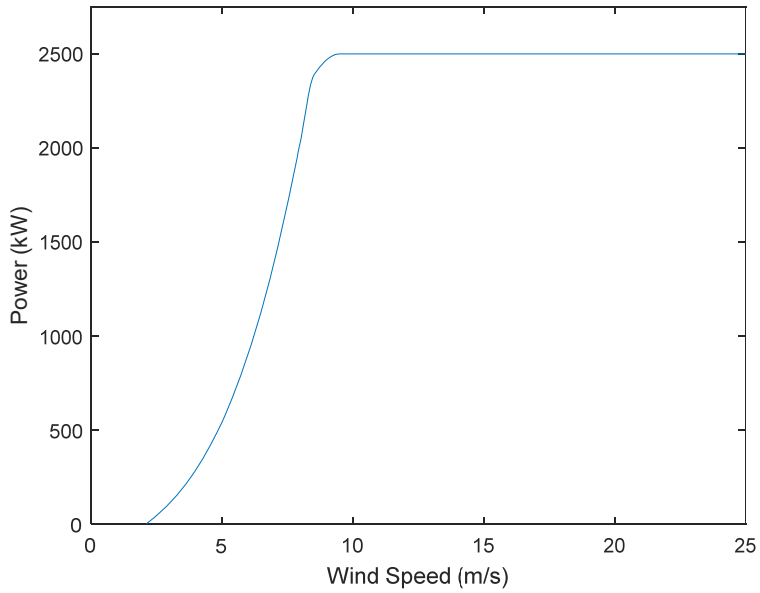


Figure 5. Case study wind turbine power vs wind speed curve.

The SCADA system of this unit is stored using 5-min sampling, and the real-time operating status parameters collected include: wind speed, spindle speed, power, voltage, current, vibration, oil pressure, and torque. In this example, operating data from 1 September–9 October 2020 was selected for analysis, with a total of 10,683 data records. The original SCADA data is shown in Table 3.

Table 3. Raw SCADA data.

Time	Average Spindle Speed (rpm)	Average Wind Direction (°)	Average Wind Speed (m/s)	Average Power (kW)	...	Cumulative Power Generation (kW·h)
1 September 2020 0:00	7.1	12.9	4	317	...	4,994,446
1 September 2020 0:05	7.7	14.1	5.1	523.5	...	4,994,491
1 September 2020 0:10	8.5	8	5.8	781.5	...	4,994,552
1 September 2020 0:15	7.1	6	4.7	397.6	...	4,994,584
1 September 2020 0:20	7.1	−10	4.1	323.6	...	4,994,609
⋮	⋮	⋮	⋮	⋮	⋮	⋮

5.2. Data Pre-Processing

5.2.1. Data Cleaning

The raw data collected by the SCADA system described the status of the equipment throughout its lifecycle, including normal operations, failures, shutdowns, maintenance, and other states. Therefore, the SCADA data contains some "dirty data", which is meaningless to the analysis and could directly or indirectly affect it. Therefore, the SCADA data requires cleaning, as follows:

1. Removal of records with status variable values that are missing or recorded as "0".
2. Referral to maintenance records to remove data recorded when the wind turbine was down for maintenance.
3. Referring to the method described in [36]: the DBSCAN-based density clustering method is used to eliminate outlier anomalies, and the truncation method is used to eliminate points where the wind speed is greater than the cut-in wind speed, but the power is still 0.

Figure 6 compares the wind speed vs power scatterplots before and after cleaning.

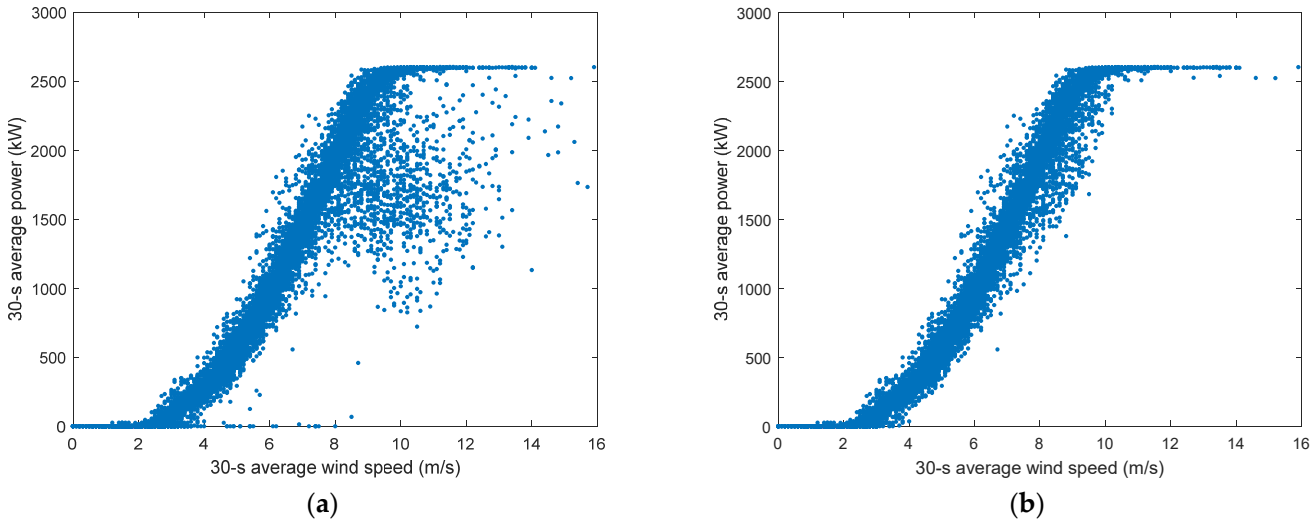


Figure 6. Scatterplots of wind speed vs. active power using (**a**) raw data and (**b**) cleaned data.

Figure 6 shows that after data cleaning, its wind speed power graph is consistent with the power curve of the wind turbine in Figure 5, and some outliers and 0-power stacking points are removed from the raw data. Hence, the data cleaning effect is achieved.

5.2.2. Data Normalization

The amount of data recorded by the SCADA system is huge, and the magnitudes of each type of operating data differ. The maximum and minimum values of different operating parameters also differ, which will affect the accuracy of the model. Therefore, the historical data needs to be normalized and mapped to the range of 0–1 to eliminate the influence of scale.

Maximum-minimum normalization is used according to the formula:

$$x_{ij}{}' = \frac{x_{ij} - \min(x_j)}{\max(x_j) - \min(x_j)} \tag{9}$$

where x_{ij} is the first i data point corresponding to variable j; $\max(x_j)$ and $\min(x_j)$ are the maximum and minimum values of variable j, respectively; and $x_{ij}{}'$ is the normalized value.

After normalization, data containing the 10 characteristic parameters of the main transmission system (selected according to the correlation analysis in Section 2) were obtained, as shown in Table 4.

Table 4. Normalized data.

Time	Average Spindle Speed	30-s Average Wind Speed	30-s Average Power	Average Main Bearing Temperature	···	Average Spindle Vibration Acceleration
1 September 2020 0:00	0.6514	0.2516	0.1217	0.7977	···	0.2165
1 September 2020 0:05	0.7064	0.3082	0.2010	0.8006	···	0.2921
1 September 2020 0:10	0.7798	0.3711	0.3	0.8064	···	0.1890
1 September 2020 0:15	0.6514	0.2956	0.1526	0.8064	···	0.2027
1 September 2020 0:20	0.6514	0.2642	0.1242	0.8064	···	0.2405
⋮	⋮	⋮	⋮	⋮	⋮	⋮

5.3. Clustering of Main Transmission System Working Conditions

Through the above analysis of the SCADA historical data, the normalized data was selected, and the main transmission system's working conditions were classified.

5.3.1. Direct Clustering of Working Conditions

To compare the effects of the working conditions before and after the method improvement, the three characteristic parameters of average spindle speed, 30-s average wind speed, and 30-s average power were selected using the *k*-means clustering algorithm to directly cluster the SCADA data of the wind turbine during normal operation, i.e., to remove data recorded during shutdown phases. The CH values of samples with different numbers of clusters from 2 to 10 were calculated, as shown in Figure 7, which shows that the CH value was largest with five clusters. Together with the shutdown phase operating conditions, there are six categories in total.

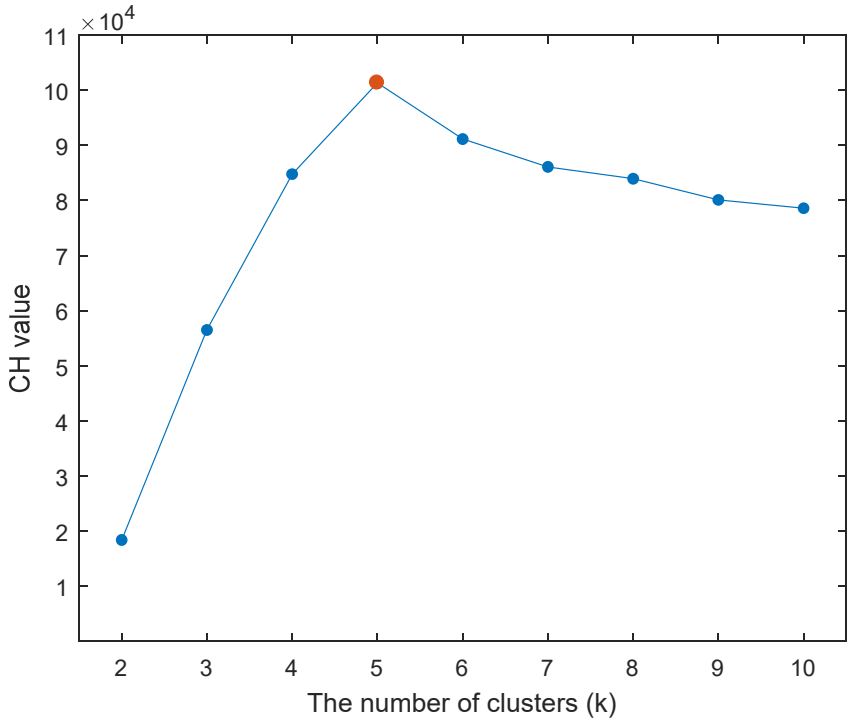


Figure 7. CH values at different number of clusters.

Figure 8 shows the results after clustering using the *k*-means clustering algorithm. It can be seen that after direct clustering of the SCADA data, the distribution of the working conditions is not obvious. The three stages of start-up, maximum wind-energy tracking, and constant speed are not clearly divided, and the working conditions are mixed.

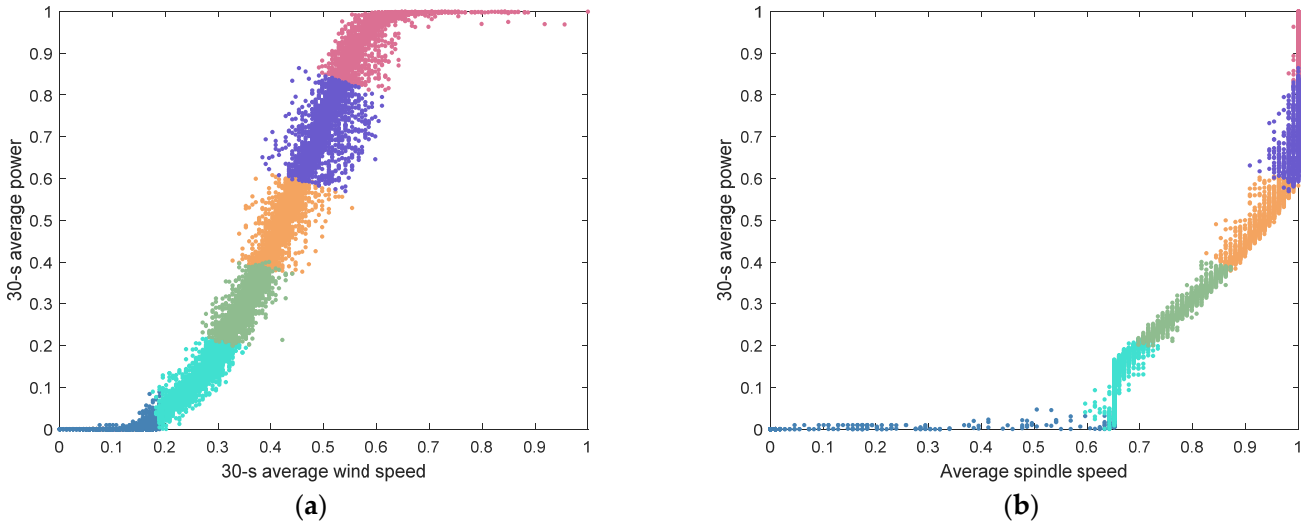


Figure 8. Direct classification of working conditions using the *k*-means clustering algorithm. (**a**) 30-s average wind speed vs. 30-s average power; (**b**) average spindle speed vs. 30-s average power.

When the turbine is shut down, the turbine does not, and its data are not considered in the condition assessment. Therefore, data from the shutdown phase (condition 1) were removed, and 2000 test samples were randomly selected from the normal operating conditions to calculate the false-alarm rate. The test samples were categorized by working condition. The working condition to which the sample belonged was determined, and the alarm threshold for each condition was determined according to the method in Section 4. The false-alarm rates were then calculated (Table 5).

Table 5. Alarm thresholds and false-alarm rates for different working conditions.

Working Condition Category (i)	Threshold (C_i)	Number of Samples Exceeding the Alarm Threshold (A_i)	Total Number of Samples (B_i)	False-Alarm Rate (R_i)
2	0.0127	27	512	5.27%
3	0.0144	21	419	5.01%
4	0.0142	11	327	3.36%
5	0.0136	15	373	4.02%
6	0.0142	6	369	1.63%
Total		80	2000	4.00%

5.3.2. Improved Working Condition Classification Method

The above analysis with the direct use of clustering algorithms improved the classification of working conditions. Firstly, according to the wind turbine output power control principle and the relationships between power and wind speed and wind turbine speed, four categories were initially defined: shutdown phase, start-up phase, maximum wind-energy tracking phase, and constant-speed phase. The results are shown in Figure 9.

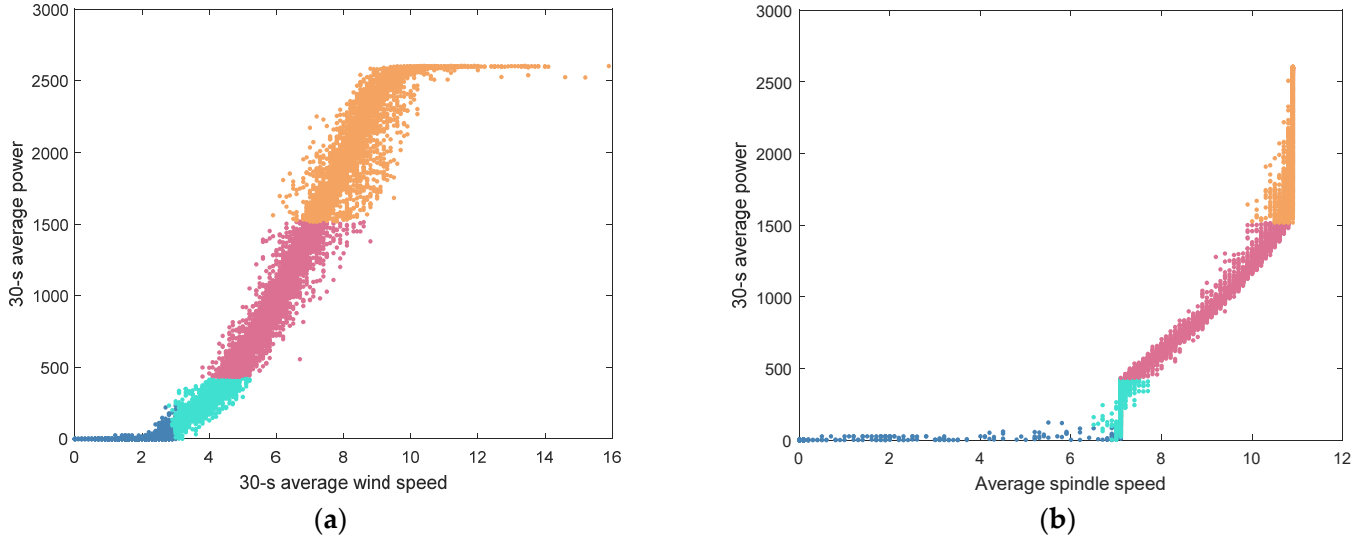
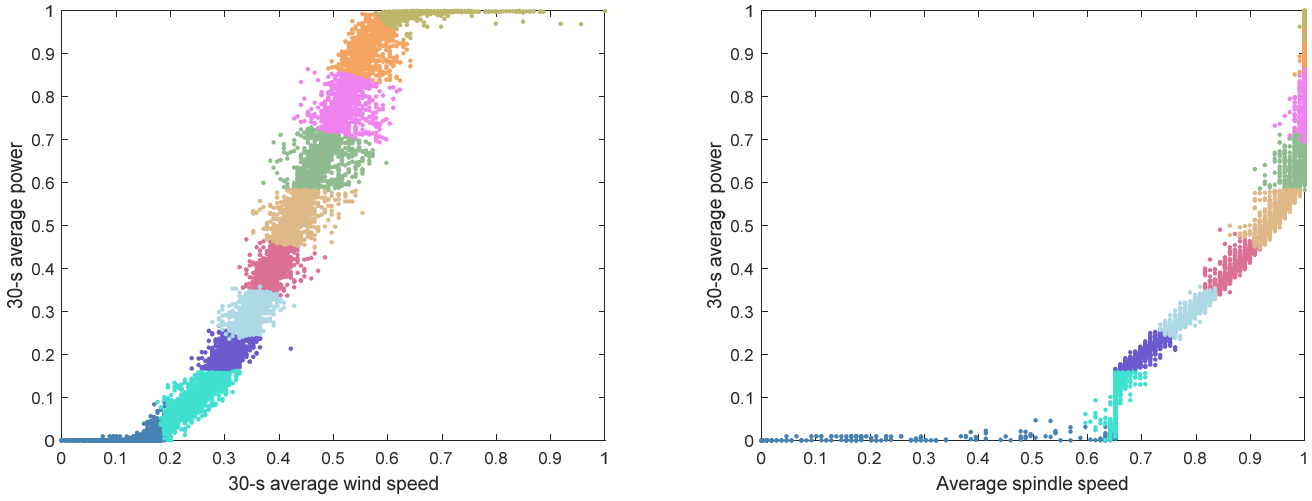


Figure 9. Preliminary working condition classification. (**a**) 30-s average wind speed vs. 30-s average power; (**b**) average spindle speed vs. 30-s average power.

After the initial classification, the third and fourth categories of conditions were subdivided using the *k*-means clustering algorithm. Firstly, for the third category, 30-s average power, 30-s average wind speed, and average spindle speed were used as the characteristic variables for clustering. The optimal number of clusters was four according to the CH criterion. For the fourth category, 30-s average power and 30-s average wind speed were used, and the optimal number of clusters was four according to the CH criterion. After aggregating the clusters, the final 10 classes of conditions were obtained, as shown in Figure 10.

To more clearly demonstrate the division of the historical working conditions, the clustered results were plotted three-dimensionally with axes representing wind speed, spindle speed, and power and are shown in Figure 11.

Figure 11 shows that, based on the analysis of operating characteristics, it is possible to divide the various phases of wind turbine operation into 10 working conditions after clustering analysis of the maximum wind-energy tracking phase and constant-speed phase. This can solve the problems of an unknown distribution of working conditions and a small number of working conditions resulting from direct clustering.



(**a**) (**b**)

Figure 10. Results after subdivision of working conditions using the *k*-means clustering algorithm. (**a**) 30-s average wind speed vs. 30-s average power; (**b**) average spindle speed vs. 30-s average power.

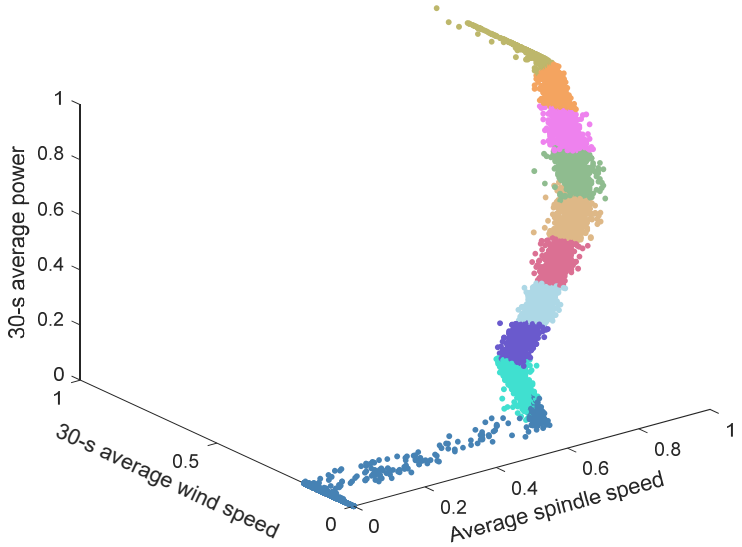


Figure 11. Three-dimensional plot of the working conditions after division.

The alarm thresholds and false-alarm rates for each condition are calculated in the same way and are shown in Table 6.

Table 6. Alarm thresholds and false-alarm rates under different operating conditions.

Working Condition Category (*i*)	Threshold (C_i)	Number of Samples Exceeding the Alarm Threshold (A_i)	Total Number of Samples (B_i)	False-Alarm Rate (R_i)
2	0.0147	7	420	1.67%
3	0.0182	3	219	1.37%
4	0.0182	0	220	0
5	0.0169	5	177	2.82%
6	0.0172	1	196	0.51%
7	0.0164	4	209	1.91%
8	0.0161	0	200	0
9	0.0159	0	200	0
10	0.0166	3	159	1.89%
Total		23	2000	1.15%

The traditional method in Figure 8 is used to classify the wind turbine operation phases into five categories, while the proposed method in Figure 11 is used to classify into nine categories. Comparing the figures, it can be seen that the various phases of the wind turbine operation can be divided based on the analysis of the wind turbine operation characteristics, and there is a clear distinction between the wind turbine start-up phase, the maximum wind energy tracking phase, and the constant speed phase. In addition, comparing the alarm thresholds and false-alarm rates derived from the 2000 test samples in Tables 5 and 6, it can be concluded that the false-alarm rates calculated directly using the *k*-means clustering algorithm basically range between 3% and 6% for each condition, with a total false-alarm rate of 4.00%. Meanwhile, the false-alarm rates derived from the proposed method are basically 0–3% for each condition, with a total false-alarm rate of 1.15%, which is an improvement compared to that of direct clustering. Therefore, the working condition classification method proposed in this paper can reasonably and effectively classify the operating conditions of a wind turbine's main transmission system and reduce the false-alarm rate in its vibration-detection process.

6. Conclusions

This paper addresses the engineering reality of the un-known real-time dynamic loads and multi-state dynamic behaviour of the main transmission system of a wind turbine, which can lead to the early onset of abnormal working conditions. A transmission system historical working condition classification method based on wind turbine operating characteristics analysis and a *k*-means clustering algorithm was proposed. The proposed method is based on the data collected using the SCADA acquisition system. This is oriented toward the division of working conditions for the later main transmission system condition assessment.

The selection method of state parameters based on correlation analysis is proposed. In order to avoid the omission of relevant state parameters and the interference of irrelevant variables to the results, the redundant variables are eliminated by using Pearson correlation coefficient. Finally, 10 state characteristic parameters of main transmission system are selected.

Most traditional working conditions are classified without taking into account the operating characteristics of the wind turbine. An improved method of working conditions classification is proposed. The working conditions are initially divided into four stages. Then, the *k*-means clustering algorithm is used to subdivide the maximum wind-energy tracking stage and constant speed stage, which account for a relatively large number of operating conditions. This provides more accurate working conditions for the subsequent condition assessment of the main transmission system.

The method was applied to a real case study of a wind turbine. To test its effectiveness, the false-alarm rates during vibration detection was calculated. Furthermore, the false-alarm rates were significantly reduced by compared with working conditions classification using the direct *k*-means clustering algorithm. The results show that the proposed classification method can solve the problems of unclear condition classification, too-few condition categories, and high false-alarm rates that could occur with direct clustering classification based on vibration detection.

The working condition classification method can be applied to the real-time monitoring of the wind turbine main transmission system at a later stage. In the process of its condition monitoring, it can reduce the false-alarm rates, improve the accuracy of main transmission system condition assessment, reduce maintenance costs, and ensure the safety and economy of wind turbines operating.

Author Contributions: Conceptualization, H.C., C.X., J.D., E.C. and J.L.; methodology, H.C. and C.X.; validation, E.C.; formal analysis, C.X.; investigation, J.L.; resources, J.D.; data curation, E.C.; writing—original draft preparation, C.X.; writing—review and editing, J.L. and H.C.; project administration, H.C.; funding acquisition, J.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of People's Republic of China (grant number 51975535 and 52075164), the Key R & D Projects of Zhejiang Province (grant number 2021C01133).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets used and/or analysed during the current study are available from the corresponding author upon reasonable request.

Conflicts of Interest: The authors declare that they have no conflict of interest.

References

1. Wang, H.; Li, X.L.; Wang, G.; Xiang, D.; Rong, Y.M. Research on failure of wind turbine gearbox and recent development of its design and manufacturing technologies. *China Mech. Eng.* **2013**, *24*, 1542–1549.
2. Dai, J.C.; Tan, Y.; Shen, X.B. Investigation of energy output in mountain wind farm using multiple-units SCADA data. *Appl. Energy* **2019**, *239*, 225–238. [CrossRef]
3. Yan, X. Research on Fault Early Warning Method for Wind Turbine Based on Condition Identification. Master's Thesis, North China Electric Power University, Baoding, China, March 2017.
4. Mei, Y.; Li, X.; Hu, Z.C.; Yao, H.; Liu, D. Identification and cleaning of wind power data methods based on control principle of wind turbine generator system. *Chin. J. Power Eng.* **2021**, *41*, 316–322.
5. Dai, J.C.; Yang, W.X.; Cao, J.W.; Liu, D.S.; Long, X. Ageing assessment of a wind turbine over time by interpreting wind farm SCADA data. *Renew. Energy* **2018**, *116*, 199–208. [CrossRef]
6. Yang, T.Y.; Zhao, L.J.; Xu, J.; Li, W.; Zhang, G.J. Abnormal identification method of wind power turbine based on Copula function. *Electr. Switchg.* **2021**, *59*, 26–31.
7. Cheng, H.F.; Zhang, Q.Y. Multi-dimensional wind turbine operating state identification based on operating data. In Proceedings of the 6th China Wind Power After-Market Communication and Cooperation Conference, Tianjin, China, 13 June 2019.
8. Ling, Y.Z. Research on Main Bearing Health Condition Assessment Method Based on Wind Turbine SCADA Data. Master's Thesis, Changsha University of Technology, Changsha, China, April 2018.
9. Gu, Y.J.; Su, L.W.; Zhong, Y.; Xu, T. An online fault early warning method for wind turbine gearbox based on operational condition division. *Electr. Power Sci. Eng.* **2014**, *30*, 1–5.
10. Dong, H.Y.; Xu, K.L.; Yang, L.X.; Gu, Y.Q.; Chen, N.N. Operational Conditions Division of Wind Turbines. In Proceedings of the 29th China Control and Decision Making Conference, Chongqing, China, 28 May 2017.
11. Xing, Y.S.; Zhuang, S.X.; Hou, Z.G.; Liao, Z.C.; Yan, W. Evaluation of wind turbine health trend based on the PCA-NAR neural network. *Electr. Autom.* **2020**, *42*, 64–66.
12. Zhang, J.; Jiang, N.; Li, H.K.; Li, N. Online health assessment of wind turbine based on operational condition recognition. *Trans. Inst. Meas. Control* **2019**, *41*, 2970–2981. [CrossRef]
13. Wang, H.; Wang, H.B.; Jiang, G.Q.; Li, J.M.; Wang, Y.L. Early Fault Detection of Wind Turbines Based on Operational Condition Clustering and Optimized Deep Belief Network Modeling. *Energies* **2019**, *12*, 984. [CrossRef]
14. Liu, W.J.; Zhao, P.F.; Zhang, W.X.; Wang, J.G. Fault diagnosis method of wind turbine based on manifold semi-supervised k-means algorithm. *Mach. Tool Hydraul.* **2020**, *48*, 191–194.
15. Jin, H.F. Research on Health Condition Monitoring of Wind Turbine Based on Data Mining. Master's Thesis, North China Electric Power University, Baoding, China, March 2018.
16. Liu, C.L.; Yan, X. Monitoring of the state of the gear box of a wind power generator unit based on the operating condition identification. *J. Eng. Therm. Energy Power* **2016**, *31*, 41–46.
17. Yin, S.; Hou, G.L.; Chi, Y.; Gong, L.J.; Hu, X.D. Prediction method for health degree of front bearing of wind turbine generator and implementation. *J. Syst. Simul.* **2021**, *33*, 1323–1333.
18. Chen, H.S.; Ma, H.Z.; Chu, X.N.; Xue, D.Y. Anomaly detection and critical attributes identification for products with multiple operating conditions based on isolation forest. *Adv. Eng. Inform.* **2020**, *46*, 101139. [CrossRef]
19. Ma, R.; Li, W.Y.; Qi, R.S. Performance degradation prognostic and health assessment using wind power data for wind turbine generation unit. *Renew. Energy Resour.* **2019**, *37*, 1252–1259.
20. Wang, D.L.; Lü, L.X.; Wang, Z.Q.; Chen, Y.; Li, X.N. Condition monitoring of wind turbine gearbox based on GMM operational condition identification and DAE. *China Meas. Test* **2021**, *47*, 89–95.

21. Han, D.P. Health Status Evaluation of Wind Turbines Based on Scada Operation Data. Master's Thesis, North China Electric Power University, Baoding, China, March 2019.
22. Han, P.P.; Xia, Y.; Ding, M.; Zhang, Y.; Lin, Z.H.; Zhu, Q.L. Equivalent modeling of wind farm based on PCA and CA-ST methods. *Acta Energ. Sol. Sin.* **2020**, *41*, 267–277.
23. Zheng, X.X.; Li, M.N.; Wang, J.; Ren, H.H.; Fu, Y. Operational conditions classification of offshore wind turbines based on kernel principal analysis optimized by PSO. *Power Syst. Prot. Control.* **2016**, *44*, 28–35.
24. Wang, F. Research on Operation Condition Classification Method for Vibration Monitoring of Wind Turbine. Master's Thesis, North China Electric Power University, Beijing, China, March 2010.
25. Ma, H.Y.; Yang, B.Y.; Peng, R.J. Research on clustering data partition algorithm based on machine learning. *Comput. Knowl. Technol.* **2021**, *17*, 9–10.
26. Putri, D.C.G.; Leu, J.-S.; Seda, P. Design of an Unsupervised Machine Learning-Based Movie Recommender System. *Symmetry* **2020**, *12*, 185. [CrossRef]
27. Zhu, X.; Zhang, S.; Jin, X.Q.; Du, Z.M. Deep learning based reference model for operational risk evaluation of screw chillers for energy efficiency. *Energy* **2020**, *213*, 118833. [CrossRef]
28. Cao, H.R.; Hou, X.G.; Feng, Y.; Bi, M. An environment modeling method in automatic layout of submarine piping. *Ship Sci. Technol.* **2020**, *42*, 103–107.
29. Bastianoni, A.; Guastaldi, E.; Barbagli, A.; Bernardinetti, S.; Zirulia, A.; Brancale, M.; Colonna, T. Multivariate Analysis Applied to Aquifer Hydrogeochemical Evaluation: A Case Study in the Coastal Significant Subterranean Water Body between "Cecina River and San Vincenzo", Tuscany (Italy). *Appl. Sci.* **2021**, *11*, 7595. [CrossRef]
30. Wang, J.H.; Jiang, J.M. Unsupervised deep clustering via adaptive GMM modeling and optimization. *Neurocomputing* **2021**, *433*, 199–211. [CrossRef]
31. Miao, K.H.; Wang, J.L.; Gao, Y.L.; Cao, C.; Xie, Y.W.; Gao, P. Robust fuzzy clustering algorithm based on adaptive neighbors. *J. Phys. Conf. Ser.* **2021**, *2025*, 012046. [CrossRef]
32. Park, J.; Jeong, J.; Park, Y. Ship Trajectory Prediction Based on Bi-LSTM Using Spectral-Clustered AIS Data. *J. Mar. Sci. Eng.* **2021**, *9*, 1037. [CrossRef]
33. Zhang, J. Data-Based Health Assessment and Fault Prediction of Wind Turbine Generator. Master's Thesis, Shanghai Jiao Tong University, Shanghai, China, February 2019.
34. Li, C.; Ye, Z.; Gao, W.; Jiang, Z. *Modern Large-Scale Wind Turbine Design Principle*, 1st ed.; Shanghai Scientific & Technical Publishers: Shanghai, China, 2013; pp. 333–341.
35. Liu, Y.Q.; Wang, F.; Shi, W.G.; Zhuo, Y. Operation condition classification method for wind turbine based on support vector machine. *Acta Energ. Sol. Sin.* **2010**, *31*, 1191–1197.
36. Wang, Y.M.; Liu, H.; Song, P.; Hu, Z.C.; Deng, X.Y.; Wu, L.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.