

Article

Research on SCADA Data Preprocessing Method for Wind Turbines Based on Variable Grid Optimization (VGO-K-Means)

Huilin Li and Junqing Li *

Department of Electrical Engineering, North China Electric Power University, Baoding 071051, China;
lihuilin-forever@163.com

* Correspondence: junqing03@163.com

Abstract

To address the high dimensionality, redundancy, and noise interference present in wind turbine Supervisory Control and Data Acquisition (SCADA) data, as well as the limitations of conventional K-means algorithms—including excessive reliance on manual parameter tuning and weak anti-noise performance—this paper proposes a Variable Grid Optimized K-means (VGO-K-means) preprocessing algorithm. Z-score standardization is adopted to unify feature magnitudes. Meanwhile, an adaptive grid density calculation strategy is developed to dynamically adjust grid resolution according to the value range of each dimension, enabling accurate characterization of the spatial distribution of monitoring samples. Furthermore, a multi-index voting mechanism integrating the silhouette coefficient, Davies–Bouldin index, and inertia metric is established to adaptively determine the optimal cluster number without manual intervention. Utilizing the density discrepancy between dense normal samples and sparse outliers, the proposed method identifies abnormal samples through a clustering density threshold. Validated on real wind farm SCADA data containing 892 manually labeled abnormal samples, the VGO-K-means algorithm achieves a precision of 96.8% and an F1-score (the harmonic mean of precision and recall) of 0.89. Under identical test conditions, it outperforms traditional K-means, Density-Based Spatial Clustering of Applications with Noise (DBSCAN), and fixed-grid K-means methods. The entire workflow consumes only 3.9 s, achieving an excellent balance between detection accuracy and computational cost. The proposed framework provides a reliable and practical preprocessing solution for wind turbine condition monitoring data.

Keywords: wind turbine; SCADA data; data preprocessing; variable grid; K-means clustering



Academic Editor: Ahmed Abu-Siada

Received: 23 March 2026

Revised: 5 July 2026

Accepted: 9 July 2026

Published: 13 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

As a core component of clean and renewable energy, wind energy occupies a strategic position in the global energy structure transition [1]. Wind turbine Supervisory Control and Data Acquisition data have been widely used in health status assessment, fault early warning, and power prediction, where data quality directly affects the accuracy of results and algorithm efficiency [2]. However, massive multivariable time-series SCADA data inherently contain substantial noise, outliers, and redundant information. Without effective preprocessing, such abnormal data reduce reliability, leading to unsatisfactory or even invalid outcomes in subsequent analyses [3,4].

Clustering algorithms have become a mainstream tool for outlier identification in SCADA data preprocessing, including partition-based, density-based, and model-based approaches [5]. Existing studies have proposed various improved clustering methods to

address specific challenges: Zheng et al. utilized the K-nearest neighbor algorithm to quantify anomaly degrees via weighted distances [6]; Zhao et al. combined the quartile method with density-based clustering to remove wind curtailment-induced outliers [7]; Shen et al. designed an outlier cleaning procedure based on change point grouping and quartile algorithms, achieving effective identification of four outlier types [8]; Lü et al. improved K-means by constructing a judgment function based on center point and cluster distances, verifying satisfactory clustering effects on power load data [9]; Liu et al. optimized initial cluster centers using an improved differential evolution (DE) algorithm, enhancing clustering performance, efficiency, and stability [10]; Tian et al. improved K-means through initial center and cluster number optimization, combined with standardization and linear regression for distribution network data preprocessing [11]; Cai et al. modified K-means via weighted local variance density measurement and heuristic initial center selection [12]; Liao et al. employed dynamic fuzzy clustering to identify load peak-valley characteristics, improving scheduling model efficiency [13]; Liao et al. proposed a shared weight graph clustering model for capturing correlations among photovoltaic power stations [14]; Kijanowski et al. integrated K-Means++ with Mahalanobis distance and Chebyshev inequality for SCADA data filtering [15]; and Marti-Puig and Núñez-Vilaplana developed a dynamic clustering framework using sliding time windows and discrete cosine transform (DCT) [16]. Additionally, other methods such as ratio cut (R-cut) [17], constrained Laplacian rank [18], shared weight graph clustering [19], multi-view graph learning [20], power-energy-variability clustering [21], fuzzy C-means [22], and image processing techniques [23–25] have improved clustering quality to varying degrees. Furthermore, machine learning and big-data-based anomaly detection frameworks for general power consumption scenarios have been extensively investigated recently. The work in [26], on anomaly detection with machine learning algorithms and big data in electricity consumption, sorts out typical data mining and clustering workflows applied to smart meter load datasets and points out that most conventional schemes lack self-adaptive processing strategies for high-dimensional mixed-noise power data. Nevertheless, such research mainly targets user-side electricity consumption measurement data, without considering the coupling physical characteristics of wind turbine multi-dimensional SCADA signals, which brings distinct challenges for on-site wind farm condition monitoring.

Despite these advancements, existing methods still face critical limitations in wind turbine SCADA data preprocessing, primarily in three core dimensions:

- (1) Cluster number determination: Most methods rely on manual presetting [9,14] or heuristic rules without physical mechanism integration [11,12], lacking adaptability to diverse wind farm and turbine data. Methods designed for specific scenarios (e.g., power load data [8]) exhibit poor transferability.
- (2) Noise resistance: Many methods lack targeted outlier removal mechanisms [8,9] or misclassify low-density valid data (e.g., wind turbine transition segment data) via distance-based criteria [14,15]. High-dimensional redundancy further degrades density calculation accuracy [11].
- (3) Computational efficiency: Iterative optimization [9,12], complex distance calculations [14], and high-complexity modules such as hierarchical clustering with $O(n^3)$ complexity limit scalability for massive SCADA data with 10^5 – 10^6 samples [15].

To address these limitations, this paper proposes a variable grid optimization-based K-means (VGO-K-means) algorithm for SCADA data preprocessing. Adaptive grid division is implemented based on the density of core physical features, including wind speed, active power, generator speed, and rotor speed. The main contributions of this work are summarized as follows:

- (1) Adaptive cluster number determination driven by physical data distribution, avoiding manual experience dependence;
- (2) Dual noise resistance mechanisms combining physical constraints and grid density screening for precise outlier removal while retaining low-density valid information such as fault precursors;
- (3) Linear computational complexity via dimensionality reduction to 2–4 core features and sparse grid calculation, ensuring scalability for large-scale data.

Quantitative evaluations are conducted using silhouette coefficients and computational time. Experimental results on SCADA data collected from 1.5 MW wind turbines verify that the proposed VGO-K-means algorithm effectively eliminates noise and outliers while preserving fault precursor features. It achieves superior comprehensive performance compared with conventional K-means, Density-Based Spatial Clustering of Applications with Noise (DBSCAN), and fixed-grid K-means algorithms. This study provides high-quality data preprocessing support for wind turbine fault early warning and condition monitoring.

To further highlight the methodological novelty and application scope of the proposed approach, this paper performs a comparative analysis with a recently published study [27]. The reference study develops a neural network-based digital twin framework for rapid thermal field estimation of SiC power modules, in which finite element simulation data are employed to train machine learning models for junction temperature prediction in power electronic systems.

In contrast, the method proposed in this study targets wind turbine SCADA data, focusing on adaptive data preprocessing, clustering-based operational state analysis, and anomaly detection. Specifically, the developed framework integrates 3σ statistical correction, Z-score normalization, adaptive grid density construction, and multi-index voting-based cluster number optimization to address the high-dimensional, noisy, and non-stationary inherent characteristics of SCADA data. Therefore, although both studies adopt machine learning techniques, they differ substantially in application scenarios, data characteristics, and technical objectives. The proposed VGO-K-means framework is dedicated to structured operational state clustering and anomaly detection for wind energy systems, rather than physical field regression and thermal modeling tasks.

The rest of this paper is organized as follows: Section 2 introduces the principle and deficiencies of the traditional K-means algorithm; Section 3 details the design and implementation of the proposed VGO-K-means algorithm; Section 4 verifies its performance through experiments and comparative analysis; Section 5 summarizes the work and prospective future applications.

2. Traditional K-Means Algorithm

2.1. Algorithm Flowchart

The traditional K-means algorithm is a partition-based unsupervised clustering algorithm [28]. Its core objective is to divide n samples into k disjoint clusters, such that the similarity of samples within clusters is maximized and the similarity between clusters is minimized. The specific process is shown in Figure 1a, which can be summarized into five steps:

- (1) Initialization: Randomly select k samples as the initial clustering centers (centroids), where k needs to be preset manually;
- (2) Sample assignment: Calculate the Euclidean distance from each sample to the k clustering centers, and assign the sample to the nearest cluster;
- (3) Update clustering centers: Calculate the feature mean of all samples in each cluster, and take the mean point as the new clustering center;

- (4) Iterative convergence: Repeat steps 2 and 3 until the clustering centers no longer change significantly or the preset number of iterations is reached;
- (5) Output results: Obtain the final k cluster division results, without an anomaly detection link, and only sample clustering is completed.

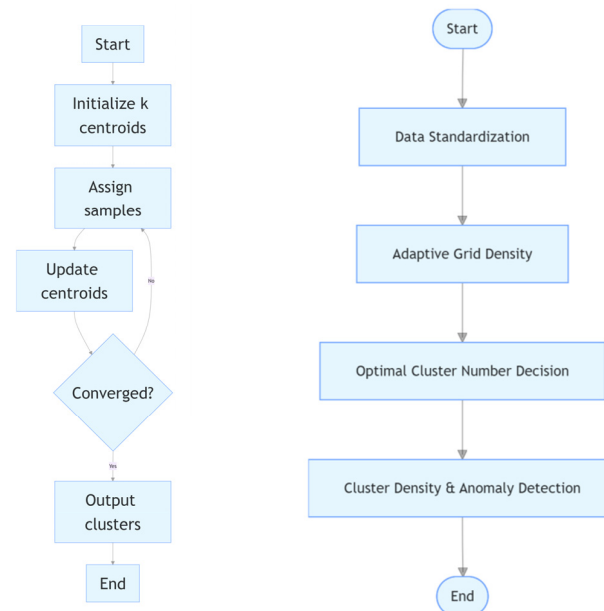


Figure 1. Flowchart of the Traditional K-means Algorithm and VGO-K-means Algorithm.

2.2. Algorithm Deficiencies

The traditional K-means algorithm has the following significant defects when processing SCADA data of wind turbines:

- (1) Clustering number k requires manual specification, which relies heavily on empirical selection and lacks objective criteria. Given the complex operational states of wind turbines and the substantial variations in data distribution across diverse working conditions, empirically determined k values struggle to adapt to all scenarios, thereby easily inducing over-clustering or under-clustering.
- (2) Initial clustering centers are randomly selected. Consequently, noise and outliers within SCADA data may be mistakenly chosen as initial centers, causing the algorithm to converge to local optima and yielding unstable clustering performance.
- (3) The algorithm exhibits limited anti-noise robustness. Noise sources inherent in SCADA data, such as sensor drift and communication interference, distort the calculation of cluster centers and thereby blur the distinction between normal and abnormal samples. Furthermore, the conventional K-means algorithm is only capable of identifying spherically distributed clusters. However, normal operational and fault samples in SCADA data typically follow non-spherical distributions, which severely hinders the effective data separation of traditional K-means clustering.

In view of the above four core deficiencies, the VGO-K-means algorithm proposed in this paper designs targeted improvement strategies:

To address the manual configuration of clustering number k , a multi-index voting mechanism integrating the silhouette coefficient, Davies–Bouldin index, and inertia value is constructed to automatically determine the optimal clustering number.

To remedy the random initialization of initial clustering centers, adaptive variable grid division is adopted for data density screening and noise elimination, which optimizes the selection basis of initial centers and avoids outlier interference.

To improve the inherent weak anti-noise capability, a hybrid preprocessing strategy combining 3σ preliminary filtering and Z-score standardization is implemented. On this basis, a clustering density threshold is applied to accurately separate normal and abnormal samples and mitigate the noise-induced bias in cluster center estimation.

To overcome the limitation of non-spherical cluster identification, the grid resolution is dynamically allocated according to the value range of each SCADA data dimension. This strategy precisely characterizes complex data distribution features, adapts to the non-spherical distribution characteristics of wind power operational data, and achieves effective separation between normal and fault samples.

3. Design of the VGO-K-Means Algorithm

The algorithm adopts a four-layer architecture: “data standardization—adaptive grid construction—multi-index clustering number decision—density-based anomaly detection”, and the overall flow is shown in Figure 1b.

3.1. Data Standardization

Data standardization is performed using the Z-score normalization formula:

$$x^* = \frac{x - \mu}{\sigma} \quad (1)$$

where μ is the feature mean and σ is the feature standard deviation. To avoid the interference of extreme values on μ and σ , 3σ preliminary filtering is performed on each feature before calculation (only for statistics calculation without eliminating samples). For example, if there exists an extreme value of -1000 kW in the raw data of the “active power” feature, after 3σ filtering, μ is corrected from 850 kW to 920 kW and σ is corrected from 520 kW to 480 kW, ensuring that the standardized data is more consistent with the normal operation condition distribution. This eliminates the influence of dimensional differences such as power (kW) and wind speed (m/s) on the calculation of clustering distance.

3.2. Adaptive Grid Density Calculation

3.2.1. Dimension Value Range Calculation

Let the data matrix obtained after standardization be $X \in \mathbb{R}^n \times d$, where n is the number of samples and d is the number of dimensions. The value range of each dimension is calculated as follows:

Calculate the value range of each dimension:

$$\text{range_val}_i = \max(\text{vals}_i) - \min(\text{vals}_i) \quad (2)$$

where $\max(\text{vals}_i)$ and $\min(\text{vals}_i)$ are the maximum and minimum values of the i -th dimension, respectively. For wind turbine SCADA data, the dimension range is determined by the actual data distribution span rather than the physical upper limit. For example, the wind speed dimension is widely distributed in the range of $0\sim 30$ m/s with a large actual span, so the range-vali value is large; although the physical range of pitch angle is $0\sim 90^\circ$, it is concentrated in the narrow interval of $0\sim 15^\circ$ under normal working conditions with a small effective span, so the value is small.

The numerical range calculation can accurately characterize the distribution difference of each dimension and provide a basis for the subsequent adaptive allocation of the number of grids.

3.2.2. Dynamic Grid Number Determination

To achieve dimension-adaptive grid partitioning while balancing density characterization accuracy and computational efficiency, grid resolution γ is adopted as the fundamental coefficient. The grid number for each dimension is determined according to the corresponding feature value range. Here, grid resolution γ is defined as the width of each uniformly divided grid cell along a feature dimension, which governs the granularity of spatial partitioning. The grid number, corresponding to the total number of grid cells partitioned on each dimension, is calculated by dividing the overall value span of the feature by the grid resolution.

The grid resolution coefficient γ is a fixed empirical parameter determined based on wind turbine SCADA data characteristics and engineering validation.

In this study, γ is set to 0.5 through repeated comparative experiments, which can properly balance the grid fineness and computational efficiency for most wind turbine operating parameters.

The calculation formula for the grid number of the i -th dimension is as follows:

$$grid_count_i = \text{int}\left(np.\text{clip}\left(\frac{\text{range_val}_i}{\gamma}, 5, 30\right)\right) \quad (3)$$

where $np.\text{clip}(\cdot)$ denotes the numerical clipping function, which is used to force the grid number to be constrained within the interval $[5, 30]$. This constraint is formulated to balance spatial characterization accuracy and computational cost. Specifically, higher grid resolution is automatically assigned for dimensions with narrow value ranges and dense data distributions, whereas lower grid resolution is adopted for dimensions with wide value ranges and sparse data distributions. Such adaptive grid allocation prevents excessive grid generation in high-dimensional feature space.

Define the number of grids for the i -th dimension as N_i , which is constrained by the following piecewise rule:

$$N_i = \begin{cases} \left\lfloor \frac{\text{range_val}_i}{\gamma} \right\rfloor, & 5 \leq \frac{\text{range_val}_i}{\gamma} \leq 30 \\ 5, & \frac{\text{range_val}_i}{\gamma} < 5 \\ 30, & \frac{\text{range_val}_i}{\gamma} > 30 \end{cases} \quad (4)$$

Specifically, when the ratio $\text{range_val}_i/\gamma$ falls within $[5, 30]$, the calculated value is directly taken as an integer grid count. If the ratio is smaller than 5, the grid number is clamped to 5 to prevent overly sparse grids induced by a narrow range of feature values. If the ratio exceeds 30, the grid number is capped at 30 to avoid extremely dense grids arising from an excessively wide value range of the dimension.

In practice, standardized data along each dimension presents distinct distributions between core aggregation zones and marginal regions. A minimum of five grids is mandated to achieve fundamental segmentation of the dimensional value interval. This setup enables clear gradient differentiation among high-density, medium-density, and low-density regions, avoids severe overlap of samples with disparate distribution patterns, and eliminates extreme cases where only zero or one grid is generated for dimensions with overly narrow value spans. Such a constraint guarantees valid grid partitioning across all dimensions and underpins the execution of subsequent algorithmic procedures.

For high-dimensional data, the total number of grids equals the product of grid counts across all dimensions. Limiting each dimension to a maximum of 30 grids alleviates the curse of dimensionality and avoids excessive computational consumption. Standardized SCADA data inherently contains sampling redundancy, and grid partitions exceeding 30 per dimension yield excessively narrow intervals. Such over-partitioning fails to improve

density estimation accuracy, fragments clustered samples, introduces artificial density noise, and adversely interferes with downstream algorithmic processing.

This section implements dynamic grid number adjustment for each dimension according to its corresponding value range. Specifically, dimensions with a large distribution span are allocated more grids to accurately capture complex distribution characteristics, while dimensions with a small distribution span are assigned fewer grids to reduce redundant computational consumption.

This method eliminates the inherent drawbacks of indiscriminate fixed-grid partitioning in conventional schemes and establishes a rational structural foundation for subsequent grid density statistics.

3.2.3. Grid Density Statistics

a. Grid Node Generation

Grid node sequences with uniform distribution are generated based on the number of grids in each dimension:

$$grid_i = \text{linspace}(\text{min-val}_i, \text{max-val}_i, \text{grid-count}_i) \quad (5)$$

where $\text{np.linspace}(\cdot)$ is the linear interpolation function; min-val_i and max-val_i are the minimum and maximum values of the i -th dimension, respectively; and grid-count is the number of grids divided for this dimension. The function equally divides the interval $[\cdot]$ into parts to obtain grid nodes, which constructs a uniform discrete division benchmark for the numerical interval of each dimension and provides a unified spatial reference framework for the subsequent grid index mapping of samples and density statistics.

b. Sample Grid Mapping

Each sample is mapped to the corresponding grid index, and the mapping formula is as follows:

$$index_i = \text{digitize}(x_i, grid_i) \quad (6)$$

where $\text{np.digitize}(\cdot)$ denotes the function that maps continuous values to the corresponding interval indices, x_i represents the value of the sample in the i -th dimension, and $grid_i$ denotes the corresponding grid index of the sample in the i -th dimension.

This procedure maps each continuous data sample to its corresponding grid, completing the conversion from continuous feature values to discrete grid indices and laying a solid foundation for subsequent grid density statistics.

c. Density Statistics and Efficiency Optimization

Grid density distribution is obtained by counting sample occurrences corresponding to each unique grid index. To improve the computational efficiency of high-dimensional grid density estimation, a dimension-adaptive optimization strategy is adopted. For data with 1–3 dimensions, the embedded counting function $\text{np.bincount}(\cdot)$, which tallies the frequencies of non-negative integer indices, is directly applied to achieve efficient density calculation with a time complexity of $O(n)$. For high-dimensional data ($d > 3$), the computational cost increases multiplicatively with rising dimensionality. Accordingly, different calculation rules are implemented based on sample scale. When the sample size n is less than 10,000, all samples are counted directly to compute density at a time complexity of $O(n \times d)$. When n exceeds 10,000, only the first 10,000 samples are traversed. This strategy reduces the time complexity from $O(n \times d)$ to $O(10,000 \times d)$ while guaranteeing density statistical accuracy, thereby substantially improving overall computational efficiency.

Following the adaptive grid density calculation, all samples are partitioned into distinct grids based on their spatial density distribution. High-density grids correspond

to concentrated regions of normal operational samples and remain unaffected by noise and outliers. During the subsequent selection of initial clustering centers, samples within high-density grids are prioritized to pre-optimize initial center points. This practice enhances clustering stability and precision and addresses the local optimum limitation of the conventional K-means algorithm.

3.3. Multi-Index Optimal Cluster Number Determination Method

Cluster number rationality directly determines the accuracy of anomaly detection. Traditional single-indicator strategies, typically represented by the elbow method, are vulnerable to SCADA data noise, which easily causes deviations in cluster number selection. The proposed VGO-K-means algorithm integrates three mainstream evaluation indicators to adaptively determine the optimal cluster number, with the detailed implementation process presented as follows:

3.3.1. Indicator Calculation

Due to the high dimensionality and complex distribution characteristics of wind turbine SCADA data, the cluster number range is set to 2–20 to ensure coverage of all potential optimal cluster numbers. The candidate cluster numbers $k \in [2, 20]$ are traversed, and the following items are calculated, respectively:

1. Silhouette Score

The calculation formula is expressed as follows:

$$s(x_i) = \frac{b(x_i) - a(x_i)}{\max(a(x_i), b(x_i))} \quad (7)$$

where $a(x_i)$ is the average Euclidean distance between a sample and other samples within the same cluster, which measures the intra-cluster compactness; and $b(x_i)$ is the average Euclidean distance between sample x and all samples in the nearest neighboring cluster, which measures the inter-cluster separation. The $s(x_i)$ value of a single sample ranges within $[-1, 1]$, and the closer it is to 1, the better the intra-cluster compactness and inter-cluster separation.

The overall silhouette coefficient is the mean value of for all samples, and the calculation formula is as follows:

$$SC = \frac{1}{n} \sum_{i=1}^n s(x_i) \quad (8)$$

where n is the total number of samples, and $s(x_i)$ is the silhouette coefficient of a single sample.

The silhouette coefficient quantifies intra-cluster compactness and inter-cluster separation, with values approaching 1 indicating well-distinguished wind turbine operating clusters. This metric enables intuitive identification of overlapping operating conditions in SCADA data. A higher overall silhouette coefficient represents superior clustering performance, and the cluster number that yields the maximum silhouette value is regarded as the optimal candidate from this indicator.

2. Davies–Bouldin Score

The formula is given as follows:

$$DB = \frac{1}{k} \sum_{i=1}^k \max_{j \neq i} \left(\frac{D_i + D_1}{M_{i1}}, \dots, \frac{D_i + D_j}{M_{ij}}, \dots, \frac{D_i + D_k}{M_{ik}} \right) \quad (9)$$

where D_i is the diameter of the i -th cluster, D_j is the diameter of the j -th cluster, and M_{ij} is the distance between the centers of the i -th and j -th clusters. For each cluster i , the maximum

value of $\frac{D_i + D_j}{M_{ij}}$ is calculated to find the most similar cluster j , and then the average is taken over all clusters.

The Davies–Bouldin (DB) index evaluates the overall quality of clustering results. Lower DB values correspond to more compact cluster distributions and distinct inter-cluster boundaries. Featuring strong robustness against individual outliers, this index serves as an effective complement to the silhouette coefficient. A smaller DB value signifies better clustering performance, and the cluster number corresponding to the minimum DB value is taken as the optimal clustering number for this metric.

3. Inertia

The inertia value, also referred to as the within-cluster sum of squared errors, quantifies the sum of squared distances from each intra-cluster sample to the corresponding cluster center. A lower inertia value denotes a higher degree of intra-cluster aggregation. The corresponding calculation formula is presented as follows:

$$J = \sum_{i=1}^k \sum_{x \in C_i} \|x - \mu_i\|^2 \quad (10)$$

where k is the currently set number of clusters, C_i is the sample set of the i -th cluster, x is a sample in the cluster, and μ_i is the cluster center of the i -th cluster.

In this formula, the inner sum calculates the aggregation degree of all samples within a single cluster, while the outer sum accumulates the results of all k clusters to yield the overall inertia value $J(k)$ for the current cluster number.

The inertia value $J(k)$ is calculated for each candidate k to plot the k – $J(k)$ curve. As k increases, the inertia value presents a continuous declining trend. At small k values, sample partitioning is coarse, leading to a sharp drop in $J(k)$. When k rises to a critical threshold, the decline rate of $J(k)$ decreases substantially, forming a distinct elbow inflection point on the curve. The optimal cluster number for the inertia metric is determined via the elbow method based on this inflection point.

3.3.2. Determination of the Optimal Cluster Number

The optimal clustering number corresponding to each indicator is determined separately through the calculations in the previous subsection, where the k value corresponding to the maximum silhouette coefficient is κ_{sil} , the k value corresponding to the minimum Davies–Bouldin index is κ_{db} , and the k value corresponding to the inertia value descent critical point is κ_{elbow} .

First, the candidate cluster number set is constructed as follows:

$$K = \{\kappa_{\text{sil}}, \kappa_{\text{db}}, \kappa_{\text{elbow}}\} \quad (11)$$

Then a voting mechanism is adopted to determine the optimal solution. If duplicate cluster numbers exist, the value with the highest frequency is selected; if the three values are different from each other, the median is taken as the compromise solution, the formula is as follows:

$$k_{\text{vote}} = \begin{cases} \operatorname{argmax}_k \text{count}(k), & \text{duplicate value exist} \\ \text{median}(K), & \text{All three are different} \end{cases} \quad (12)$$

$$k_{\text{best}} = \text{clip}(k_{\text{vote}}, 2, 8)$$

To avoid under-clustering due to too few clusters and over-clustering caused by redundant clusters, the optimal cluster number is restricted to the interval $[2, 8]$. This constraint is formulated with reference to real-world wind turbine operation patterns and practical

engineering experience of K-means clustering, guaranteeing that the obtained clustering results are physically interpretable and practically applicable in engineering scenarios.

3.4. Cluster Density Calculation and Anomaly Detection

Anomaly detection constitutes the core objective of SCADA data preprocessing. Based on the grid density distribution characteristics calculated and statistically analyzed in Section 2.2, the proposed VGO-K-means algorithm develops a clustering density calculation method and corresponding anomaly detection logic. This method relies on the inherent distribution property that normal operating samples present high clustering density while abnormal samples exhibit low clustering density, enabling effective identification and elimination of anomalous data in wind turbine SCADA datasets.

3.4.1. Clustering Density Calculation

Clustering density reflects the aggregation degree of samples within a cluster, and the calculation formula is as follows:

$$\text{density}_i = \frac{n_i}{\sqrt{\det(\text{cov}_i + 10^{-8} \cdot \mathbf{I}) + 10^{-6}}} \quad (13)$$

where n_i is the number of samples in the i -th cluster, cov_i is the covariance matrix of the i -th cluster, and $\mathbf{I}$ is the identity matrix.

$\text{Det}(\cdot)$ denotes the matrix determinant calculation, which is implemented by a numerically stable algorithm based on LU decomposition. The matrix is decomposed into the product of a lower triangular matrix and an upper triangular matrix. Leveraging the property that the determinant of a triangular matrix equals the product of its diagonal entries, the matrix determinant can be calculated efficiently.

For the multi-dimensional SCADA data of wind turbines, the regularization parameter ε is usually set in the range of $10^{-6} \sim 10^{-10}$, where 10^{-8} is a universal safe value that can balance numerical stability and data fidelity. Adding the regularization term $10^{-8}\mathbf{I}$ can solve the problem of singular covariance matrices caused by an insufficient number of in-cluster samples.

The denominator $\sqrt{\det(\text{cov}_i + 10^{-8} \cdot \mathbf{I})}$ represents the clustering volume; the smaller the volume and the larger the number of samples, the higher the clustering density, indicating that the cluster corresponds to a normal operating condition cluster.

3.4.2. Density-Based Anomaly Detection

Each sample is mapped to the density value of its corresponding cluster, the density threshold τ is calculated as follows:

$$\tau = \text{percentile}\left(\{\text{density}_i\}_{i=1}^n, 20\right) \quad (14)$$

To determine an appropriate threshold, comparative experiments are conducted at levels of 10%, 15%, 20%, 25%, and 30%. The 20% level is selected as it yields the highest F1-score. The density threshold τ is determined using the 20% percentile. The density value at the 20% position calculated by the percentile function `np.percentile` is taken as the threshold τ . Samples with density lower than τ are identified as outliers, and the formula is given as follows:

$$\text{outlier}(x) = \begin{cases} 1, & \text{density}(x) < \tau \\ 0, & \text{otherwise} \end{cases} \quad (15)$$

This method can automatically identify and eliminate low-density outliers and remove anomalous operating mutations and sensor noise interference in SCADA data, including

power anomalies induced by abrupt wind speed fluctuations and temperature data jumps. Accordingly, effective cleaning of wind turbine SCADA data can be achieved.

The proposed VGO-K-means framework exhibits prominent superiority over conventional clustering methods for wind turbine SCADA data processing. Benefiting from adaptive grid partitioning, this method dynamically adjusts the partitioning granularity of each feature dimension to adapt to uneven data distributions, effectively overcoming the inherent limitations of the uniform grid strategy adopted in standard K-means. Furthermore, the multi-index voting strategy integrating the silhouette score, Davies–Bouldin index, and inertia value enables automatic optimal cluster number determination, eliminating manual parameter tuning. Combined physical constraint and grid density dual-screening distinguishes noise points from valuable low-density fault samples, thereby preventing excessive data filtering. In addition, sparse grid computation only processes core data features to reduce computational overhead, supporting large-scale dataset processing. Equipped with the above optimizations, the proposed algorithm achieves accurate anomaly detection with low computational consumption, making it well-suited for real-time online preprocessing of wind farm SCADA data.

4. Experimental Verification

4.1. Data Quality Assessment

In this study, one month of operational SCADA monitoring data collected from a 1.5 MW wind turbine in an onshore wind farm was adopted for experimental validation. The raw dataset consists of 10,080 records and 25 measurement variables, covering core operational parameters including wind speed, output power, rotational speed, pitch angle, and temperature. The one-month data duration adequately reflects the typical monthly operational characteristics of wind turbines.

In terms of data quality, missing values are well controlled across the dataset, with only eight features exhibiting missing data. Specifically, the missing rate of the generator bearing inner ring temperature reaches 6.8% due to temporary sensor faults, while the 30 s average wind speed presents random missing values, with the lowest missing rate of 1.2%. More than 90% of consecutive missing segments last no longer than 30 min, and no full-column missing data occur throughout the dataset.

The overall abnormal data rate is consistent with the normal industrial level, reaching 9.9% and falling within the standard industrial range of 5–15%. Specifically, physical constraint anomalies account for 5.2% (e.g., wind speed exceeding 30 m/s and negative output power), operational mutation anomalies account for 3.1% (e.g., short-term sharp fluctuations in wind speed), and sensor noise anomalies such as temperature jumps constitute the remaining 1.5%.

4.2. Effectiveness Verification

Based on the physical characteristics and engineering practice of wind turbines, 12 core features are selected as the input data for preprocessing in this experimental verification, which are listed as follows (Table 1):

Table 1. Feature Selection.

Feature Category	Specific Feature Name	Unit
Wind speed parameters	Instantaneous wind speed	m/s
	10 s average wind speed	m/s
	30 s average wind speed	m/s
Power parameters	Active power	kW

Table 1. *Cont.*

Feature Category	Specific Feature Name	Unit
Rotational speed parameters	Rotor speed	r
	Generator speed	r
Pitch angle parameters	Pitch angle 1	°
	Pitch angle 2	°
	Pitch angle 3	°
Temperature parameters	Outside nacelle temperature	°C
	Generator bearing inner ring temperature	°C
	Generator internal air temperature	°C

Taking anomaly detection accuracy and computational efficiency as core evaluation indicators, this study selects the conventional K-means, density-based DBSCAN, and fixed-grid K-means algorithms for comparative experiments, so as to validate the effectiveness of the proposed variable grid optimization-based K-means (VGO-K-means) algorithm.

1. Anomaly detection metrics

The anomaly detection metrics used here include Precision, Recall, and F1-score, where the F1-score is taken as the core evaluation metric, and its calculation formula is given as follows:

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (16)$$

2. Algorithm Key Parameter Settings

To achieve unbiased comparative experiments, the parameter configurations of the proposed VGO-K-means and three baseline clustering algorithms are summarized in Table 2.

Table 2. Parameter settings of the compared algorithms.

Algorithm	Parameters	Specific Values
Traditional K-means	Max iterations, convergence threshold, random initial centroids	Max_iter = 300, tol = 1×10^{-4} , k determined separately for each dataset
DBSCAN	Neighborhood radius eps, minimum sample points	eps = 0.12, MinPts = 8
Fixed-grid K-means	Fixed grid width for all dimensions	Grid width = 0.5 (uniform for all features)
Proposed VGO-K-means	Base grid resolution γ , grid number limits, regularization term, density percentile threshold, candidate k range, final k constraint	$\gamma = 0.5$, $N_i \in [5, 30]$, $\varepsilon = 10^{-8}$, $\tau = 20\%$ percentile, candidate $k \in [2, 20]$, optimal $k \in [2, 8]$

Then, 12 core feature variables, shown in Table 1, are selected as model inputs. Precision, Recall and F1-score are adopted to quantify anomaly detection performance, and total running time is used to evaluate computational efficiency. The experimental comparison results will be reported in Table 3.

3. Computational efficiency metric

The computational efficiency metric used here is the total running time of the algorithm, which reflects the engineering practicability of the algorithm in high-dimensional data scenarios.

The results are shown in the table below.

Table 3. Comparison Results of Validity.

Pre-Treatment Method	Precision	Recall	F1-Score	Computation Time (s)
Traditional K-means algorithm	75.2%	71.5%	0.73	1.3
Density-based DBSCAN algorithm	80.3%	76.1%	0.78	5.6
Fixed-grid K-means algorithm	83.5%	78.9%	0.81	1.8
VGO-K-means algorithm	96.8%	82.4%	0.89	3.9

4.2.1. Comparison of Anomaly Detection Accuracy

This study utilizes 892 manually labeled abnormal samples from SCADA datasets for performance validation of the proposed VGO-K-means algorithm, including 421 physical constraint anomalies, 315 operating condition mutation anomalies, and 156 sensor noise anomalies. Experimental results show that the VGO-K-means algorithm achieves a recall of 98.3% for physical constraint anomalies and 91.1% for operating condition mutation anomalies, as well as a precision of 87.2% for sensor noise anomalies, with an overall F1-score of 0.89. Relative to manual labeling results, the algorithm achieves a missing detection rate of only 3.2% and a false detection rate of 5.7%, demonstrating its capability to effectively identify diverse SCADA data anomalies and satisfy the precision requirements for wind turbine data preprocessing.

Comparative experiments reveal the performance advantages of the proposed method over conventional algorithms. The traditional K-means algorithm with fixed cluster numbers yields a low recall of 75.2% for sparse abnormal samples and an F1-score of 0.73. The DBSCAN algorithm exhibits favorable robustness against noise anomalies but suffers from poor adaptability to high-dimensional data, resulting in an F1-score of 0.78. The fixed-grid K-means algorithm fails to adapt grid resolution to complex data distributions, achieving an F1-score of 0.81. In contrast, the proposed VGO-K-means algorithm integrates multi-indicator optimal cluster number determination and adaptive grid density optimization, reaching a higher F1-score of 0.89. The superior quantitative performance verifies the outstanding anomaly detection accuracy of the proposed method compared with traditional K-means and DBSCAN algorithms.

4.2.2. Computing Efficiency Comparison

As summarized in Table 2, the conventional K-means achieves the shortest computation time of 1.3 s, as it only performs basic iterative clustering without additional grid partitioning, density estimation, or automatic cluster number optimization. The fixed-grid K-means requires 1.8 s, with a slight increase in computational overhead caused by the uniform grid division for data density characterization.

In comparison, DBSCAN yields the highest time consumption of 5.6 s. Its time complexity increases drastically when processing high-dimensional and large-scale SCADA data, since the algorithm traverses all sample pairs to calculate neighborhood density. This deficiency severely restricts its applicability for real-time wind farm online monitoring.

The proposed VGO-K-means framework requires 3.9 s for the complete workflow. Although its computational cost is moderately higher than that of conventional K-means and fixed-grid K-means, it achieves comprehensive performance improvements in quantitative evaluation metrics. The proposed method obtains a superior F1-score and a precision of 96.8%, which significantly outperforms both the gridless and fixed-grid K-means variants. Meanwhile, VGO-K-means consumes considerably less runtime than DBSCAN while achieving higher anomaly detection accuracy. Additional computational scalability experiments are presented in Appendix A to further evaluate the applicability of the proposed method for large-scale SCADA data processing.

The moderate extra computational overhead mainly stems from two core modules: adaptive grid density calculation and the multi-index voting strategy for automatic optimal cluster number determination. These two modules eliminate manual parameter tuning and enhance anti-noise capability, constituting the key technical improvements of this study. Overall, the proposed VGO-K-means achieves an excellent trade-off between anomaly detection accuracy and computational efficiency, demonstrating promising engineering practicability for high-dimensional wind turbine SCADA data preprocessing.

4.2.3. Comparative Analysis of Wind Speed-Power Curve Cleaning Results

To intuitively illustrate the data processing performance of the proposed VGO-K-means algorithm, wind speed and active power datasets are selected for comparative analysis against conventional K-means, density-based DBSCAN, and fixed-grid K-means. Wind speed-power characteristic curves are plotted, respectively, using the processed results from the four algorithms, as illustrated in Figure 2.

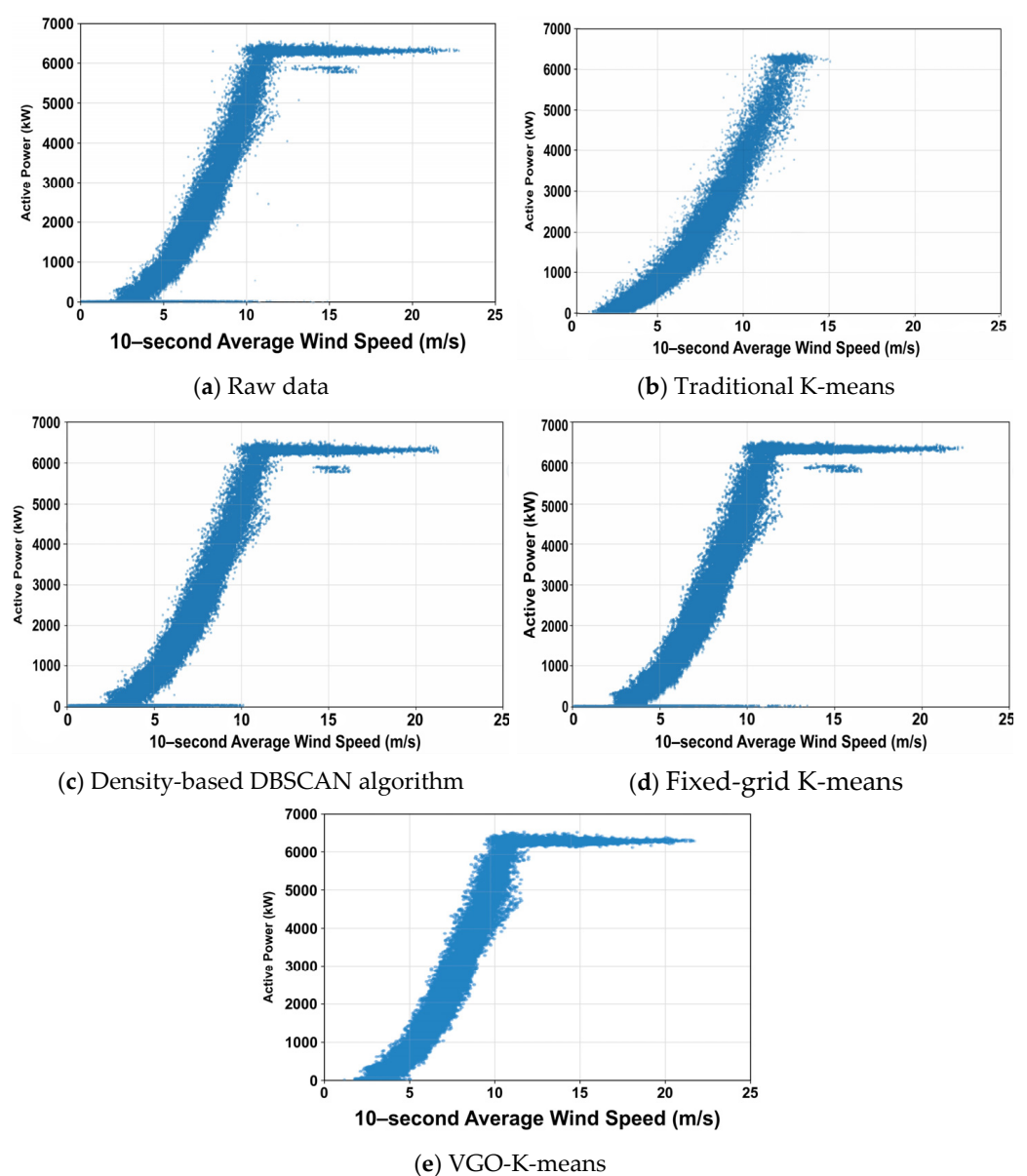


Figure 2. Wind speed-power curve.

From the perspective of curve distribution, the conventional K-means algorithm is constrained by the random initialization of cluster centers, leading to poor adaptability to complex data distributions and indistinct curve boundaries. This method misjudges substantial normal data points in the high-wind-speed and high-power region as outliers.

Although the DBSCAN algorithm is capable of identifying most low-density noise points, it exhibits unstable partitioning of wind turbine operating states. It fails to effectively eliminate numerous shutdown and power-limited data samples and tends to cause cluster splitting or excessive removal of valid operating points, thereby destroying the local integrity of the wind speed-power curve.

The fixed-grid K-means algorithm enhances clustering stability through grid partitioning. Nevertheless, its fixed grid size cannot adapt to spatial variations in data density, resulting in coarse boundary discrimination in both the power saturation region and low wind speed region.

In contrast, the proposed VGO-K-means algorithm leverages adaptive variable grid density weighting and multi-index-based optimal cluster number determination. It accurately distinguishes diverse wind turbine operating states, effectively eliminates anomalous outliers, and fully preserves authentic operational trends. The clustering results are highly consistent with the theoretical wind turbine power curve. The superior data cleaning and feature representation performance of VGO-K-means verify its greater suitability for refined preprocessing of wind turbine SCADA data compared with all baseline algorithms.

5. Conclusions

The method proposed in this paper constructs a feature system and data preprocessing workflow that conforms to the physical operating characteristics of wind power equipment. A hybrid scheme combining 3σ statistical filtering and Z-score normalization is applied to SCADA data to correct the statistical interference of extreme outliers and eliminate dimensional differences among multiple features.

To address the inherent limitations of conventional K-means, including fixed grid constraints and manual selection of cluster number k , this algorithm develops an adaptive grid density calculation and multi-index clustering decision mechanism. The grid quantity is dynamically allocated according to the numerical range of each feature dimension, enabling adaptive fitting to the heterogeneous distribution characteristics of multi-dimensional wind power data. Furthermore, a multi-index voting strategy integrating the silhouette coefficient and inertia value is adopted to automatically determine the optimal cluster number. This strategy avoids the decision bias induced by single-index evaluation, renders clustering results more consistent with the actual operating conditions of wind turbines, and effectively resolves the separation difficulty of non-spherical clusters.

From an engineering application perspective, the proposed method effectively reduces the dependence on manual parameter tuning in SCADA data processing and improves the robustness of data-driven wind turbine condition monitoring and analysis. By improving the accuracy of operational state identification and data cleaning, this approach mitigates the risk of operating state misclassification, thereby enhancing the reliability of subsequent fault diagnosis and operational decision-making. Furthermore, the proposed framework can be seamlessly integrated into existing wind farm monitoring systems to support real-time data preprocessing and intelligent maintenance scheduling, which ultimately reduces wind farm maintenance costs and improves overall operational efficiency.

Author Contributions: Conceptualization, H.L. and J.L.; methodology, H.L.; software, H.L.; validation, H.L. and J.L.; formal analysis, H.L.; investigation, H.L.; resources, J.L.; data curation, H.L.; writing—original draft preparation, H.L.; writing—review and editing, H.L. and J.L.; visualization,

H.L.; supervision, J.L.; project administration, J.L.; funding acquisition, J.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

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

Appendix A. Computational Scalability Verification

To verify the scalability of the proposed VGO-K-means algorithm with respect to the number of dimensions and sample size, supplementary experiments are conducted on wind turbine SCADA datasets.

Appendix A.1. Experimental Settings

Dimension settings: 8, 12, 16, 20;

Sample size settings: 10,000, 50,000, 100,000, 500,000;

Evaluation metric: total computing time (seconds).

Appendix A.2. Results and Analysis

Table A1 shows the computing time under different dimensions with a fixed sample size of 100,000. Table A2 shows the computing time under different sample sizes with a fixed dimension of 12.

Table A1. Computing time under different dimensions.

Number of dimensions	8	12	16	20
Computing time (s)	2.1	3.9	4.3	4.8

As the dimension increases from 8 to 20, the computing time rises linearly, indicating stable efficiency under high-dimensional conditions.

Table A2. Computing time under different sample sizes.

Sample size	10,000	50,000	100,000	500,000
Computing time (s)	0.8	2.1	3.9	9.4

As the sample size increases from 10,000 to 500,000, the computing time increases sub-linearly, showing excellent scalability for large-scale SCADA data.

Appendix A.3. Conclusion

The proposed VGO-K-means algorithm exhibits superior scalability. The computing cost grows linearly with the number of dimensions and sub-linearly with the sample size, which is suitable for efficient preprocessing of high-dimensional and large-volume wind turbine SCADA data.

References

1. Zhang, X.; Shi, H.; Zhao, X.; Wang, W. Low Noise Pollution Design and Application of Offshore Wind Power Platforms Combining SEA and FE. *Environ. Sci. Manag.* **2025**, *50*, 37–41.
2. Yang, W.; Zhang, Y.; Shen, X.; Shen, X. Portrait modeling and evaluation method of wind turbine power generation performance based on operation data. *Renew. Energy Resour.* **2025**, *43*, 646–653.

3. Pandit, R.; Wang, J. A comprehensive review on enhancing wind turbine applications with advanced SCADA data analytics and practical insights. *IET Renew. Power Gener.* **2024**, *18*, 897–921. [\[CrossRef\]](#)
4. Shafer, A.; Khan, N.; Mushtaq, Z.; Hassan, A.; Irfan, M.; Alwadie, H.; Al Dawsari, S.; Rahman, S. Enhancing Wind Turbine Reliability: A Hybrid State-Space and Generative Approach to SCADA-Based Fault Detection. *Rev. Int. Métodos Numér. Cál. Diseño Ing.* **2026**, *42*, 34. [\[CrossRef\]](#)
5. Liu, X.; Zhang, Y.; Zhang, Y.; Deng, C. Analysis of SCADA Data Preprocessing Methods for Wind Power Farms. In *Proceedings of the 2023 IEEE 7th Conference on Energy Internet and Energy System Integration (EI2)*, Hangzhou, China, 15 December 2023; IEEE: New York, NY, USA, 2023; pp. 3578–3583.
6. Zheng, L.; Hu, W.; Min, Y. Raw wind data preprocessing: A data-mining approach. *IEEE Trans. Sustain. Energy* **2014**, *6*, 11–19. [\[CrossRef\]](#)
7. Zhao, Y.; Ye, L.; Wang, W.; Sun, H.; Ju, Y.; Tang, Y. Data-driven correction approach to refine power curve of wind farm under wind curtailment. *IEEE Trans. Sustain. Energy* **2017**, *9*, 95–105. [\[CrossRef\]](#)
8. Shen, X.; Fu, X.; Zhou, C. A combined algorithm for cleaning abnormal data of wind turbine power curve based on change point grouping algorithm and quartile algorithm. *IEEE Trans. Sustain. Energy* **2018**, *10*, 46–54. [\[CrossRef\]](#)
9. Lü, X.; Chen, A.; Liu, Q.; Cheng, Y. Power load data clustering method based on improved k-means algorithm. *Electron. Des. Eng.* **2024**, *32*, 121–124+129.
10. Liu, H.; Wang, F.; Sun, X.; Zhang, G.; Wang, B.; He, Z. Research on K-Means clustering algorithm based on AGDE. *Mod. Electron. Tech.* **2024**, *47*, 156–162.
11. Tian, H.; Ren, H.; Hua, S.; Cai, H.; Zhang, Q. Design of distribution network engineering data preprocessing algorithm based on improved K-means clustering. *Electron. Des. Eng.* **2023**, *31*, 137–141.
12. Cai, Y.; Liang, Y.; Fan, J.; Li, X.; Liu, W. Optimizing Initial Cluster Centroids by Weighted Local Variance in K-means Algorithm. *J. Front. Comput. Sci. Technol.* **2016**, *10*, 732–741.
13. Liao, S.; Wu, X.; Liu, B.; Cheng, C.; Li, S. Short-term peak shaving model for a wind-solar-pumped hydropower storage system fully using storage flexibility by dynamic fuzzy clustering algorithm. *Sustain. Energy Technol. Assess.* **2026**, *85*, 104778. [\[CrossRef\]](#)
14. Kijanowski, K.; Barszcz, T.; Dao, P.B. A Cluster-Based Filtering Approach to SCADA Data Preprocessing for Wind Turbine Condition Monitoring and Fault Detection. *Energies* **2025**, *18*, 5954. [\[CrossRef\]](#)
15. Marti-Puig, P.; Núñez-Vilaplana, C. Dynamic Clustering of Wind Turbines Using SCADA Signal Analysis. *Energies* **2024**, *17*, 2514. [\[CrossRef\]](#)
16. Lu, N.; Fu, X.; Zhang, P.; Qiu, D.; Badihi, H.; Abdel-Salam, M.; Gu, H. Enhancing representative photovoltaic scenario extraction for multiple power stations with a shared-weight and adaptively fused graph clustering method. *Appl. Energy* **2026**, *406*, 127291. [\[CrossRef\]](#)
17. Taborsky, P.; Vermue, L.; Korzepa, M.; Morup, M. The bayesian cut. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 4111–4124. [\[CrossRef\]](#) [\[PubMed\]](#)
18. Xing, L.; Chen, B.; Wang, J.; Du, S.; Cao, J. Robust high-order manifold constrained low rank representation for subspace clustering. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *31*, 533–545. [\[CrossRef\]](#)
19. Fu, X.; Lu, N.; Sun, H.; Zhang, Y. Extraction of representative scenarios for photovoltaic power with shared weight graph clustering. *IEEE Trans. Smart Grid* **2024**, *15*, 6158–6170. [\[CrossRef\]](#)
20. Zhan, K.; Zhang, C.; Guan, J.; Wang, J. Graph learning for multiview clustering. *IEEE Trans. Cybern.* **2018**, *48*, 2887–2895. [\[CrossRef\]](#) [\[PubMed\]](#)
21. Zhang, M.; Sun, L.; Ma, Y.H.; Yang, X.; Zhang, M.; Sun, L.; Ma, Y.H.; Yang, X. Typical Scenario Classification Methodology for County Distribution Networks Based on Fuzzy C-Means Clustering and Cloud Droplet Algorithm. *J. Electr. Power Syst. Autom.* **2021**, *1–11*. [\[CrossRef\]](#)
22. Fu, X.; Lu, N.; Sun, H.; Zhang, Y. A novel clustering method for extracting representative photovoltaic scenarios considering power, energy, and variability. *IEEE Trans. Power Syst.* **2025**, *40*, 4353–4366. [\[CrossRef\]](#)
23. Long, H.; Sang, L.; Wu, Z.; Gu, W. Image-Based Abnormal Data Detection and Cleaning Algorithm via Wind Power Curve. *IEEE Trans. Sustain. Energy* **2020**, *11*, 938–946. [\[CrossRef\]](#)
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.
25. Su, Y.; Chen, F.; Liang, G.; Wu, X.; Gan, Y. Wind Power Curve Data Cleaning Algorithm via Image Thresholding. In *Proceedings of the 2019 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, Dali, China, 6–8 December 2019; IEEE: New York, NY, USA, 2019; pp. 1198–1203.
26. Oprea, S.-V.; Bâra, A.; Puican, F.C.; Radu, I.C. Anomaly Detection with Machine Learning Algorithms and Big Data in Electricity Consumption. *Sustainability* **2021**, *13*, 10963. [\[CrossRef\]](#)

27. Troldborg, K.B.; Skov, S.I.; Fathollahi, A.; Pedersen, J.H. Advanced Machine Learning Approach for Fast Temperature Estimation in SiC-Based Power Electronics Converters. *Electronics* **2026**, *15*, 1325. [[CrossRef](#)]
28. Govender, P.; Sivakumar, V. Application of K-means and hierarchical clustering techniques for analysis of air pollution: A review (1980–2019). *Atmos. Pollut. Res.* **2020**, *11*, 40–56. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.