Article

Data Processing Method of Mine Wind Speed Monitoring Based on an Improved Fuzzy C-Means Clustering Algorithm

Wei Zhang, Yucheng Li * and Junqiao Li

School of Safety and Emergency Management Engineering, Taiyuan University of Technology, Taiyuan 030002, China
* Correspondence: liyucheng@tyut.edu.cn; Tel.: +86-18306895838

Abstract: Analyzing and processing mine wind speed monitoring data is the key to realizing intelligent ventilation and real-time calculation of the ventilation network. According to the characteristics of the artificial regulation of a mine ventilation system, a local regression fuzzy C clustering algorithm is proposed in this paper, which combines local outlier processing with global air volume state analysis. Firstly, the algorithm uses the robust local weighted regression principle to analyze and preprocess the data locally, determines the risk degree of the abnormal data according to the identified times of outliers, determines the clustering number according to the clustering validity function, and analyzes the global air volume fluctuation according to the clustering results. The results show that most outliers are identified in data preprocessing. Still, the processing of dense outliers is weak, related to the window width setting and weighting multiple. The number of clusters can represent the fluctuation of the ventilation state and the pre-processed cluster centers are 4.4% lower than the original data because most of the outliers are higher than the average data. According to the law of air volume balance, the clustering results can pave the way for the global deduction of mine wind speed. There is an implicit relationship between data preprocessing and the clustering process, and when intensive outliers are not eliminated, they may be identified as separate clusters. The research of this paper points out the direction of mine wind speed data analysis, which can provide a theoretical basis for intelligent mine ventilation and real-time calculation of the ventilation network.

Keywords: wind speed anomalies; local regression; fuzzy clustering; cluster validity; wind state

1. Introduction

1.1. Research Motivation

With the rapid application and popularization of artificial intelligence technology in various working fields worldwide, safety and intelligence will become a new mode for coal enterprises to improve industry competitiveness and sustainable development as one of the subsystems of intelligent mine construction [1]; the mine ventilation system puts forward higher requirements for the accurate calculation of airflow, stability of air network, reliable decision-making of disaster, and so on [2]. The mine ventilation system is still in the artificial or semi-artificial stage, far behind the open, intelligent design. It is imperative to break through the common problems of the industry and improve the intellectual level of mine ventilation [3]. The analysis and processing of mine wind speed monitoring data is the key to controlling ventilation state, real-time calculation of ventilation network, accident diagnosis, system optimization, and so on. A ventilation network is a dynamic balance system that is affected by mining, transportation, personnel activities, geological conditions, ventilation system regulation, and other factors; the data of wind speed sensors will have an abrupt short-term phenomenon [4]. It will also show long-term changes in the adjustment process of ventilation facilities [5]. To realize intelligent ventilation and improve the accuracy of ventilation network calculation, a data processing method is needed to eliminate the noise in the mine wind speed data and mine itself and analyze the long-term change information in the data.
1.2. Related Work

The anomaly processing methods of large sensor data samples can be divided into the data fusion method, signal processing method, and multi-intelligence body fusion method [6,7]. The data fusion method aims to infer abnormal values by combining the average values of some machine learning algorithms and multiple sensors [8,9]. The idea of the signal processing method is to restore the signal to form the initial sample and then use the principal element analysis, wavelet analysis, and other methods to put forward the signal features [10]. A multi-agent for data processing method treats various processing methods as agents, gives the decision structure and route after fusion, and then judges whether the data are abnormal [11,12]. Experts in specific fields (such as wind fields) have carried out much research on analyzing and predicting wind speed data, and most deal with their nonlinear characteristics based on time series [13,14]. For example, Altan [15] et al. aiming at the non-stationarity and randomness of wind speed data, developed a new hybrid 20 WSF model based on long-term memory (LSTM) network, 21 decomposition methods, and grey wolf optimizer (GWO). A hybrid wind speed prediction model is proposed, considering both the accuracy and stability of wind speed prediction. Chun-Ying Wu [16] used fully integrated empirical mode decomposition (CEEMD) to divide the original wind speed series into a set of intrinsic mode functions and then applies the extreme learning machine (ELM) optimized by multi-objective grey wolf optimization (MOGWO) to achieve excellent prediction performance. Karasu [17] used nonlinear autoregressive exogenous (NARX) neural networks to estimate the relationship between some parameters and wind speed. Based on the new combination method of two-stage data preprocessing technology, three-component prediction model, and multi-objective optimization algorithm, Ying Wang [18] proposed a unique combination forecasting system, which can decompose and reshape the original data to reduce noise and chaotic interference. Ying Nie [19] offered a two-way wind speed prediction and analysis system, which realizes the dual calculation of wind speed determination point prediction and interval prediction.

Different from the characteristics of wind speed in other engineering fields, the change of mine wind speed is controlled by man as a whole, and its change trend changes with the evolution of ventilation facility control, and the information on ventilation facility variation may be implied in the wind speed data. Therefore, the analysis and processing of wind speed data are more critical [20]. The interpretation of mine wind speed can be divided into two categories: random variables and process variables, in which random variables can be understood as noise data in a local sense, including facility drop, train and personnel passage, abnormal detection equipment, and so on. The analysis of the uncommon degree of noise data can reveal the safety problem of randomness. The process variable has a definite change trend, which generally lasts for a specific time, including shaft extension, structure damage, roadway penetration, cage lift, air door installation, ventilation power regulation, and so on. The analysis of process variables is of global significance. The extraction and analysis of the characteristics of different periods, mining the ventilation state change information and comparing with the facility regulation measures, find the system’s hidden danger.

For data information mining in the global sense, most of the clustering methods are based on the data characteristics [21]. Many studies have used the mean C fuzzy clustering (FCM) algorithm. Mean C fuzzy clustering is an algorithm for clustering according to data characteristics. At present, there are many improved fuzzy C-means clustering algorithms. Yaxiong Chi [22] proposed a large-scale GRN model based on FCM in 2019 and obtained the FCM algorithm’s limitation: its state value must be normalized to $[0, +\infty]$, which does not meet the requirements of $[0, +\infty]$ required by the model. Zhou Jin [23] proposed to use advanced meta-heuristic methods and hybrid optimization techniques with fuzzy logic to optimize the objective function of clustering. The centralized clustering problem is solved by cooperating only with neighboring peers in a distributed pattern at each peer. Mudan Li [24] selected an effective wind speed, rotor speed, pitch angle, and output power which reflected the operation characteristics of the wind turbine as several groups of indicators,
and weighted the samples based on traditional FCM, which can more accurately reflect the dynamic characteristics of wind farm access points. Most of the above methods pay attention to the computational aspects of FCM itself, such as initial value and objective function, and do not solve the problems in engineering applications. What needs to be solved urgently in the field of mine ventilation is to determine the clustering number, which represents the number of changes in wind speed in the overall state, and also indicates the hidden dangers such as the damage to structures and the decrease in power energy consumption, which can be used for safety personnel to inspect and repair the facilities.

1.3. Necessity of Research Based on Challenges of the Literature

After summarizing the above literature, combined with the engineering characteristics in the field of mine ventilation, the following conclusions are drawn:

1. Most traditional wind speed analysis is based on random prediction. In the field of mine ventilation, the wind speed is artificially controlled as a whole, therefore the study of wind speed should focus on the identification and location of noise data and analyze the degree of variation to provide a theoretical basis for mine ventilation workers to investigate safety hidden dangers;

2. Although researchers have carried out much research on the calculation accuracy and speed of clustering in theory, in the engineering practice of mine ventilation, the clustering number means the overall fluctuation of air volume. Therefore, the most important thing is to select a reasonable method to determine the clustering number;

3. In the field of mine ventilation, there is a specific relationship between random noise and the fluctuation of the overall state of air volume, therefore it is necessary to combine the two for analysis, not only to determine the location of the noise but also to obtain the information of the overall fluctuation of air volume and explain the possible implicit relationship between the two.

1.4. Novelty and Main Contributions

Based on the shortcomings of the above work and methods, this paper puts forward a method of wind speed data analysis according to the characteristics of mine ventilation systems in engineering practice and outputs. It analyzes the results, which provides theoretical support for mine intelligence. The main innovation and contribution of this paper can be summarized in the following four points:

1. According to the demand for mine intelligent construction and the engineering practice of mine ventilation, the wind speed processing methods in other engineering fields are compared and the analysis idea of mine wind speed data is put forward, which combines the local outliers with the overall air volume fluctuation;

2. The robust local regression method is used to identify the preprocessing wind speed data, identify outliers, locate the abrupt data, and classify its risk for mine workers;

3. The preprocessed wind speed data are clustered by fuzzy C-means clustering and the clustering validity function is introduced. The clustering number is determined through the analysis of separation degree and compactness and the corresponding ventilation state is analyzed;

4. The clustering results after data preprocessing are compared with the origin. The clustering results and the implicit relationship between noise data and clustering results are analyzed, which provides a theoretical basis for the further integration of the two.
2. Principle of FCM Clustering Algorithm Based on Local Regression Feature

2.1. Robust Local Regression

The outliers in the data set refer to a small part of the point that deviates from the trend of most data [25]. The recognition and processing of outliers is the basis for the overall fuzzy clustering of data sets. It is equated with the local anomaly recognition and processing of data sets [26,27]. This process can also be called data preprocessing.

First, several pieces of data are divided into small intervals and the regression weights are calculated for each data point in the gap. The following function gives the weight:

$$w_i = \left(1 - \left(\frac{|x - x_i|}{d(x)}\right)^3\right)^3$$ (1)

In the formula: $x$ is the value that needs to be smooth; $x_i$ is the $i$th value on both sides of $x$; and $d(x)$ is the two norms of interval length (also known as window length). A weighted polynomial fits the samples in the interval to obtain the smooth value of $x$.

In the second step, to enhance the robustness of the data, $MAD$ is used to give the data robust weight in the fitting process to eliminate the outliers. $MAD = median(|r|)$, the median of the absolute value of the deviation between the data point and the sample median. The bi-square function gives the weight:

$$w_i = \begin{cases} 
1 - \left(\frac{r_i}{6MAD}\right)^2, & |r_i| < 6MAD \\
0, & |r_i| > 6MAD
\end{cases}$$ (2)

In the formula: $r_i$ are the residuals for the $I$ data point generated by the smoothing process. If $r_i < 6MAD$, the robust weight is 0 and the point is excluded from the smoothing calculation.

By repeating the above two processes, the double smooth curves of regression and robustness can be obtained.

Finally, the center of these regression curves is connected to obtain a complete regression curve.

2.2. Cluster Validity Function

The Xie–Beni index uses compactness to evaluate the aggregation degree within the class and uses dispersion to assess the isolation degree between classes. The Xie–Beni index transforms the problem of evaluating the effectiveness of fuzzy clustering into a situation of solving the optimal clustering number [28]. The essence is to set different clustering numbers and obtain the Xie–Beni index value and determine the optimal clustering number through the Xie–Beni index value. The calculation formulas of intra-class compactness $Var(U, c)$ and inter-class separation $Sep(U, c)$ are as follows:

$$Var(U, c) = \sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij} ||x_j - c_i||^2$$ (3)

$$Sep(U, c) = \min ||c_i - c_j||^2, i, j \in \{1, 2, \cdots, c\}, i \neq j$$ (4)

$$S(U, c) = \frac{\sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij} ||x_j - c_i||^2}{\min ||c_i - c_j||^2}$$ (5)

In the formula: $c_i, c_j$—the clustering center of the $i, j$ class; $x_j$—the $j$th sample; $u_{ij}$—the degree of membership; and the $j$-th sample belongs to category $i$. 
The smaller the distance between the center of the sample and the cluster to which it belongs, and the larger the distance to the center of different groups, the more favorable the fuzzy clustering division is. In terms of the Xie–Beni index, the compactness coefficient should be as small as possible to achieve the best clustering effect and the separation coefficient $Sep(U, c)$ should be as significant as possible. Therefore, when evaluating the final clustering effect, the smaller the Xie–Beni index is, the better the clustering effect is.

2.3. FCM Clustering Algorithm

FCM clustering algorithm is based on a specific objective function; it divides the set $X_i$ of data monitored during the $T_i$ cycle of the wind speed sensor into $c$ classes. Assuming that the algorithm sample $x_j$ belongs to class $i$ with the degree of membership $u_{ij}$, the objective function and constraints are as follows:

$$J = \min \left( \sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij}^m \|x_j - c_i\|^2 \right)$$

subject to $\sum_{i=1}^{c} u_{ij} = 1, j = 1, 2 \ldots, n$

In the formula:

- $J$—objective function;
- $c_i$—Class $i$ sample data center;
- $m$—membership factor, which represents the sample’s degree of ease, is generally 2;
- $x_j - c_i$—Euclidean distance from the sample $x_j$ to the center $c_i$.

Solving the above equation, the Lagrange multiplier method transforms the constrained optimization into the unconstrained optimization problem [29]. Then, let the partial derivatives of $u_{ij}$, $c_i$, $\lambda_j$ and other variables in the function be 0. Finally, the iterative formulas of variables $u_{ij}$ and $c_i$ are as follows:

$$u_{ij} = \frac{1}{\sum_{k=1}^{C} \left( \frac{\|x_i - c_k\|}{\|x_i - c_k\|} \right)^{2/m}}$$

$$c_i = \frac{\sum_{i=1}^{N} u_{ij}^m \cdot x_j}{\sum_{i=1}^{N} u_{ij}^m}$$

2.4. Process of Local Regression FCM Algorithm

The algorithm flow is shown in Figure 1. Local regression FCM algorithm determines the location and frequency of abnormal wind speed values, selects the cluster number by Xie–Beni index after smoothing the original data, and determines each cluster’s cluster center and sample number. The local regression FCM algorithm meets the requirements of intelligent ventilation; determining the position and number of abnormal values is equivalent to choosing the position and number of abnormal wind speed values, which staff can use to check sensor faults. Smoothing the original data is equal to reducing the noise of the wind speed data set based on maintaining the data characteristics so that it can participate in the ventilation network solution; the determination of the number of clusters is equivalent to the conclusion of the number of ventilation state fluctuations; the cluster center and the sample number in the class belong to different fluctuation states.
3. Results

3.1. Data Sources

To verify the practicability of the algorithm, the thermal anemometer is used to measure the wind speed in the simulated mine roadway in the laboratory. Every 5 s, the wind speed is measured and recorded, a total of 300 groups of data. The experimental data are shown in Table 1. Figure 2 shows the four kinds of equipment mainly used in the experiment. (a) is a simulated mine roadway, which is used to simulate the complex roadway network of a real mine; (b) is the different obstacles in the pipeline, the obstacles can change the wind area, change the wind pressure of the branches, and then change the wind speed of the measuring point, which is used to simulate the overall fluctuation of air volume over a period of time; (c) is the high-precision wind speed monitoring instrument used in the experimental process; and (d) is a fan that provides initial ventilation power. The outliers in the wind speed data set are affected by sprinkling paper near the probe of the wind speed sensor.
3. Results

3.1. Data Sources

To verify the practicability of the algorithm, the thermal anemometer is used to measure the wind speed in the simulated mine roadway in the laboratory. Every 5 seconds, the wind speed is measured and recorded, a total of 300 groups of data.

Table 1. Monitoring wind speed sensor data in a certain period in coal mines.

| No. | Values/(m/s) | No. | Values/(m/s) | No. | Values/(m/s) | No. | Values/(m/s) | No. | Values/(m/s) |
|-----|--------------|-----|--------------|-----|--------------|-----|--------------|-----|--------------|
| (1) | 2.62         | (111)| 2.42         | (171)| 2.53         | (231)| 2.54         | (491)| 1.21         |
| (2) | 2.42         | (112)| 2.49         | (172)| 2.34         | (232)| 2.45         | (492)| 1.01         |
| (3) | 2.51         | (113)| 2.39         | (173)| 2.66         | (233)| 2.67         | (493)| 1.40         |
| (4) | 2.37         | (114)| 2.64         | (174)| 2.65         | (234)| 2.30         | (494)| 1.29         |
| (5) | 2.54         | (115)| 2.38         | (175)| 2.63         | (235)| 2.48         | (495)| 1.38         |
| (6) | 2.41         | (116)| 2.39         | (176)| 2.40         | (236)| 2.47         | (496)| 1.38         |
| (7) | 2.56         | (117)| 2.37         | (177)| 2.54         | (237)| 2.48         | (497)| 1.37         |
| (8) | 2.58         | (118)| 2.39         | (178)| 2.31         | (238)| 2.61         | (498)| 1.32         |
| (9) | 2.60         | (119)| 2.47         | (179)| 2.47         | (239)| 2.43         | (499)| 1.29         |
|     |              |     |              |     |              |     |              |     |              |
| (10)|              |     |              |     |              |     |              |     |              |
| (11)|              |     |              |     |              |     |              |     |              |
| (12)|              |     |              |     |              |     |              |     |              |
| (13)|              |     |              |     |              |     |              |     |              |
| (14)|              |     |              |     |              |     |              |     |              |
| (15)|              |     |              |     |              |     |              |     |              |

Table 2 shows the data of risks and values of each risk level calculated according to the method described in the previous section.

Figure 2. Diagram of the experimental equipment: (a) Simulated mine roadway; (b) Obstacles in the roadway; (c) Thermal anemometer; (d) Variable frequency fan.

3.2. Data Preprocessing Results

Based on the principle of locally weighted robust regression, the original data were pre-processed and the interval length of the data set was chosen as seven based on previous experience. The results are shown in Figure 3.

This process identifies and processes local anomalies in the raw data. It accomplishes two main tasks: one is to remove outliers and the other is to smooth the data noise while maintaining the original data fluctuation trend. The window length in the local regression is seven, meaning that the window traverses the entire data set with each piece of data maintaining the original data fluctuation trend. The window length in the local regression is seven, meaning that the window traverses the entire data set with each piece of data maintaining the original data fluctuation trend. The window length in the local regression is seven, meaning that the window traverses the entire data set with each piece of data maintaining the original data fluctuation trend.

In total, 81 pieces of data were identified as outliers. In total, 81 pieces of data were identified as outliers. In total, 81 pieces of data were identified as outliers. In total, 81 pieces of data were identified as outliers. In total, 81 pieces of data were identified as outliers. In total, 81 pieces of data were identified as outliers. In total, 81 pieces of data were identified as outliers. In total, 81 pieces of data were identified as outliers.

Hazard analysis tables for data $n \geq 3$ are shown in Table 2 below.
3.2. Data Preprocessing Results

Based on the principle of locally weighted robust regression, the original data were preprocessed and the interval length of the data set was chosen as seven based on previous experience. The results are shown in Figure 3.

![Figure 3. Outliers Detection and Processing Results.](image)

Table 2. Risk table of monitoring data.

| No.  | Values/(m/s) | Times | Risk  | No.  | Values/(m/s) | Times | Risk  |
|------|--------------|-------|-------|------|--------------|-------|-------|
| 26   | 3.03         | 3     | low   | 442  | 1.10         | 3     | low   |
| 72   | 2.93         | 3     |       | 107  | 3.22         | 4     |       |
| 134  | 3.08         | 3     |       | 296  | 3.09         | 4     |       |
| 180  | 2.43         | 3     |       | 461  | 1.11         | 4     | medium|
| 181  | 0.46         | 3     |       | 466  | 1.29         | 4     |       |
| 208  | 0.51         | 3     | low   | 114  | 2.64         | 5     |       |
| 209  | 2.48         | 3     |       | 462  | 1.01         | 5     |       |
| 252  | 2.67         | 3     |       | 392  | 3.29         | 6     |       |
| 261  | 2.47         | 3     |       | 481  | 2.61         | 6     |       |
| 266  | 2.68         | 3     |       | 47   | 4.35         | 7     | high  |
| 267  | 2.38         | 3     |       | 274  | 3.16         | 7     |       |
| 301  | 5.21         | 3     |       | 373  | 3.20         | 7     |       |
| 441  | 4.71         | 3     |       | —    | —            | —     | —     |

Comparing the results in Table 2 with Figure 3, the evaluation table provides an objective assessment of the risk of the data. High-risk outliers occurred five times or 1% of the total data; medium-risk outliers occurred six times or 1.2% of the entire data, and low-risk happened 14 times or 2.8% of the whole data.

Sample 284 is the only one identified as an abnormal value, which is not listed in Table 2. This is because the outlier detection algorithm in this step depends on the change in the surrounding data. The data before and after the 284th sample have large fluctuations, which makes the MAD value larger and the robustness weight larger, therefore they cannot be recognized as an outlier.

Constant abnormal values were detected at samples 180 and 181, 208 and 209, 441, and 442, and the three groups of abnormal values appeared at the fluctuation of the overall wind speed state. For example, in samples 180 and 181, the wind speed jumps from a lower state to a higher state and the algorithm identifies it as an abnormal value at the junction.
3.3. Wind Speed State Fluctuation Results

After data preprocessing, it is necessary to analyze the implied wind speed state fluctuation information from the global perspective. According to the intra-class compactness $\text{Var}(U,c)$, inter-class separation $\text{Sep}(U,c)$ and the Xie–Beni index, the number of wind speed state fluctuations is finally determined and the results are shown in Figure 4.

![Figure 4. Compactness versus cluster number.](image)

As the number of clusters increases, the compactness tends to decrease. When the number of clusters is three the degree of change of compactness is the largest, and then it remains a gentle downward trend. It shows that when the number of clusters increases from two to three, the effect is the most obvious from the compactness level. When the number of clusters continues to grow from three, the compactness decreases slightly and the compactness has no significant effect on the clustering effect.

The overall trend of separation decreases as the number of clusters increases. When the number of clusters is two, three, and four, it can be considered that the separation between different classes is more effective and there is a clear boundary between classes. When the number of clusters is more significant than four, the degree of separation is small. It changes slowly, indicating that the boundaries between categories are unclear and that the clustering effect is ineffective.

When the number of clusters is three, the Xie–Beni index value is the smallest. That is to say, a cluster number of three is the best.

Taken together, the optimal number of clusters is three. At a cluster number of three, the original data were clustered using the local regression FCM algorithm and the FCM algorithm, respectively, and the results were as follows:

As seen in Figure 5a, the local regression FCM algorithm membership graph is smoother and the difference between different classes is more prominent. As seen in Figure 5b, the FCM membership graph shows the number of mutations related to the data set’s outliers. When solving practical engineering problems, local regression FCM membership graph plots can be more intuitive in analyzing the problem.

As can be seen from Figure 6, the local regression FCM algorithm divides the data into three categories. The first class has a clustering center of 4.979, which contains samples 302 to 442; the second class has a clustering center of 2.497, which includes samples 1 to 180, 209 to 301, and 442; and the third class has a clustering center of 0.998, which contains samples 181 to 208 and 443 to 500.
Figure 5. Comparison of the degree of membership graphs of the two methods: (a) Local regression FCM; (b) FCM.

Figure 6. Comparison of the clustering centers of the two methods.
This represents a certain number of fluctuations in wind speed around the three clustering centers of 4.979, 2.497, and 0.998 during this period. Samples belonging to each category do not necessarily appear in a concentrated manner, e.g., samples 181–208 and 443–500 are in the same class at the same time but occur in a very different order. This requires a judgment analysis of the samples in each cluster with the membership graph.

The clustering centers obtained by the FCM algorithm are 4.986, 2.514, and 0.988, respectively, and the local regression FCM algorithm results are 4.4% lower than the FCM results. This is due to the outliers that increase the clustering centers during the iterations. These outliers are not average data and should be removed. Substituting the outliers into the calculation will affect the final result.

All data are then rearranged by category to make a scatter diagram, as shown in Figure 7. The local regression FCM algorithm eliminates the outliers in the second clustering, which makes the clustering results more clear. There are some outliers at the junction of different classes, such as sample 141, which is determined by the algorithm characteristics of local regression FCM. When the calculation window passes through the junction, weighted regression smoothes an intermediate value at the center of the two categories.

Figure 7. Comparison of the sample clustering plots of the two methods.

4. Discussion

In this section, the results and errors of the algorithm are analyzed, compared with the wind speed processing in other fields, and the algorithm’s applicability in mine ventilation engineering practice is examined.

4.1. Data Preprocessing

In the first stage of the operation of the regression fuzzy C clustering algorithm, starting from the local meaning of the wind speed, the outliers are also identified and processed while keeping the changing trend of the data so that it can better participate in the second stage operation. At the jump of air volume, there is an anomaly in the identification and treatment of outliers, and the sudden change between classes will identify the air volume at the jump as continuous outliers. Because the characteristic of the algorithm tends to reduce the abnormal trend, after the smooth calculation of the algorithm, new outliers are added at the connection of two different ventilation states. Because the calculated value of MAD is too large for the dense outliers, some abnormal data are not identified as outliers and can not be eliminated smoothly.

Compared with the data cleaning algorithm based on machine learning and random matrix theory, the denoising algorithm in this paper solves the problem of robustness by setting different window widths and weighted calculations [30,31]. The wind speed data of mine ventilation also need to be used to solve the real-time ventilation network, therefore the calculation speed is higher. If more time is spent in the data preprocessing...
stage, it is difficult to achieve the function of real-time calculation. Machine learning needs many samples to train the model but the mines in actual production are often complex and changeable, which is impractical.

The setting of window width and weighted multiple is essential for identifying outliers. The selection of window width reflects the smoothing ability of the algorithm. If the value is too large, there will be transition smoothing and the ability to maintain the characteristics of the original data will be reduced; if it is too small, the outliers can not be well removed. The weighted multiple represents the criteria for determining outliers [32]. Setting these too high or too low will affect the recognition results. The setting of these two main factors has good flexibility and can be set according to the different data characteristics of other mine locations. In addition to the need for further analysis of the sensitivity of the algorithm parameters, it is also necessary to make it have the ability of hierarchical noise reduction. Some sensors are located in a core position and their noise reduction algorithms can sacrifice other features to compensate for their accuracy. However, the role of some sensors is not core and their noise reduction algorithms can offer accuracy to pay for other characteristics.

4.2. Fuzzy Clustering

The optimal number of clusters is affected by the initial value, the number of iterations and the initial value selection is random. Within the limited number of iterations, the compactness, separation, and Xie–Beni index values will have a slight deviation but the trend remains unchanged. Before clustering, a reasonable range of initial values should be given to accelerate the convergence speed and improve the algorithm's accuracy. The global significance of the local regression FCM algorithm can be seen from the clustering results. From Figure 6, according to the amplitude of wind speed fluctuation, the wind speed set in Figure 6 can be divided into five categories, in which the serial data numbers are 1: 180, 181: 208, 209: 301, 302: 442 and 443: 500, respectively. The five data types fluctuate around the three clustering centers of 4.979, 2.497, and 0.998, and the degree of discretization among the data is significant. According to the calculation results of the algorithm and combined with the regulation means of underground ventilation facilities, managers can investigate and evaluate the hidden dangers in the ventilation system.

Determining the mine ventilation state according to the clustering number benefits the global deduction of the monitoring data. The wind speed value has significant time variation characteristics. The wind network structure, system components (roadways, air doors, wind windows, wind bridges, main fans, local fans, etc.), and characteristic parameters (air volume, air resistance, atmospheric state parameters, working air volume, and wind pressure of main fans, etc.) are not fixed values, but dynamic. The complete air volume data should make the monitoring air volume consistent with the empirical value and follow the law of air volume distribution. Although the on-site technicians can accurately grasp the air volume of some branches and have a general understanding of the air volume of most of the other branches, it is not easy to control the air volume distribution pattern of the whole ventilation network. There are two kinds of methods to obtain the target air volume of the ventilation network. The first method is to determine the air volume of the roadway comprehensively and the second method is to select part of the air volume data as the test standard. The method of comprehensive measurement of roadway air volume has the following disadvantages: (1) large workload; (2) when the ventilation system is abnormal, not only the characteristics of the components are difficult to obtain, but also the amount of data obtained is limited; (3) there are errors between the obtained data, which do not fully meet the node equation. It is challenging to eliminate errors artificially. When selecting some branches to preprocess the wind speed data and carry on the cluster analysis, according to the calculation results, comparing the air volume changes of each branch according to the air volume distribution law, we can roughly calculate the air volume changes of the remaining branches.

Compared with the fuzzy clustering methods in the image and medical field, the clustering validity function is introduced in this paper, which solves the problem of cluster
number certainty very well [33,34]. The data dimension in the field of mine ventilation is single but it implies the relevant information of ventilation state, therefore it is necessary to analyze the interpretability of clustering. Clustering based on a fuzzy decision tree combines the flexibility of fuzzy division with the interpretability of the decision tree, which can be added to wind speed clustering in subsequent research [35].

4.3. Analysis of the Relationship between Preprocessing and Clustering Results

In general, data preprocessing serves for clustering, and the smoothed data set makes the clustering results more accurate and closer to the actual clustering center. In the case of this paper, the value of the clustering center of the preprocessed wind speed data is smaller than that of the original data because the outliers of the original data are mostly higher than the average data. The appearance of continuous outliers may indicate the overall fluctuation of the air volume state.

There may be errors in the clustering process in two cases: one is that when the wind volume fluctuates as a whole, at the connection of the two, data preprocessing often generates new outliers in the smoothing process, and second, when outliers appear densely, some abnormal data are not identified as outliers and can not be eliminated smoothly. When these outliers appear densely near a specific value, the algorithm will identify them as a single cluster and list them separately, which will be confused with average wind speed clustering. This is related to the robustness of data preprocessing and clustering validity function. Therefore, it is necessary to set the smoothing process at the connection of the air volume state separately to reduce the generation of new outliers and identify the intensive outliers in the data set to enhance the algorithm’s applicability.

5. Conclusions and Future Works

The main results are as follows: (1) the analysis and processing of mine wind speed data should be combined with local noise analysis and global air volume fluctuation. Based on the principles of local nonparametric optimization and fuzzy clustering calculation, the regression fuzzy C clustering algorithm can identify and deal with the local outliers and global outliers of underground wind speed. The algorithm comprehensively considers the location and times of outliers and the determination and analysis of air volume state, which makes establishing the model more reasonable; (2) the results of data preprocessing show that high, medium, and low-risk outliers account for 1%, 1.2%, and 2.8% of the data set, respectively. The setting of window width and weighted multiple is the critical factor affecting outlier identification. Regarding the abnormal global state, the local regression FCM algorithm determines that the number of clusters is three through the clustering validity function. Due to avoiding the influence of outliers, the clustering center is 4.4% lower than that of the FCM algorithm. Finally, the air volume state is determined to fluctuate around 4.979, 2.497, and 0.998 centers; (3) there is an implicit relationship between clustering results and local outliers, and the appearance of continuous outliers indicates the change in the overall state of air volume. When the data preprocessing creates an error in identifying outliers, it will increase the number of clusters, and outliers will be identified as one class separately, which is related to the robustness of the preprocessing algorithm.

Future research will be carried out from the following aspects:

1. We will analyze the data anomalies caused by different kinds of random variables, find the differences and relationships between them, and make a risk classification comparison table of various random variables, which provides a theoretical basis for checking the hidden dangers of the mine ventilation systems;
2. The uncertainty and sensitivity of parameters such as window width and weighted multiple in data preprocessing will be analyzed to solve the problems of data preprocessing failure and clustering errors in extreme cases;
3. According to the clustering results of this paper, the monitoring data of the mine will be deduced globally according to the law of air volume balance, which provides a theoretical basis for the intelligent ventilation of the mine.
Author Contributions: Conceptualization, W.Z. and Y.L.; methodology, W.Z.; software, J.L.; validation, W.Z., Y.L. and J.L.; formal analysis, W.Z.; investigation, Y.L.; resources, Y.L.; data curation, J.L.; writing—original draft preparation, W.Z.; writing—review and editing, W.Z.; visualization, J.L.; supervision, Y.L.; project administration, W.Z.; funding acquisition, Y.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

References

1. Zhang, Y.F.; Chen, B.L.; Liu, X.; Tang, Y. Path planning for an intelligent coal mine monitoring system based on artificial neural network. Sens. World 2016, 10, 31–33.
2. Cheng, J.; Yang, S. Data mining applications in evaluating mine ventilation system. Saf. Sci. 2012, 50, 918–922. [CrossRef]
3. Yan, Z.; Wang, Y.; Fan, J. Research on Safety Subregion Partition Method and Characterization for Coal Mine Ventilation System. Math. Probl. Eng. 2021, 2021, 1–11. [CrossRef]
4. Dong, L.; Sun, D.; Han, G.; Li, X.; Hu, Q.; Shu, L. Velocity-free Localization of Autonomous Driverless Vehicles in Underground Intelligent Mines. IEEE Trans. Veh. Technol. 2020, 69, 9292–9303. [CrossRef]
5. Deng, Q.; Liu, H.; He, Y.; Huang, T.; Zhou, H.; Bao, Y. Uniformity and Energy Evaluation of Equal Cross-section Ventilation System (ECVS) for Long Tunnel in Underground buildings. Energy Built Environ. 2022, 3, 86–94. [CrossRef]
6. Arai, K.; Seto, K. Data fusion method for Earth Observation Satellite Data based on Wavelet MRA. Dtsch. Med. Wochenschr. 2006, 22, 233–236.
7. Xu, Z.; Hu, Q.; Ehsani, M. Estimation of Effective Wind Speed for Fixed-Speed Wind Turbines Based on Frequency Domain Data Fusion. IEEE Trans. Sustain. Energy 2012, 3, 57–64. [CrossRef]
8. Ghareshchopogh, F.S.; Namazi, M.; Ebrahimi, L.; Abdollahzadeh, B. Advances in Sparrow Search Algorithm: A Comprehensive Survey. Arch. Comput. Methods Eng. 2022. [CrossRef]
9. Wang, S.; Wang, J.; Lu, H.; Zhao, W. A novel combined model for wind speed prediction—Combination of Linear Model, Shallow Neural Networks, and Deep learning Approaches. Energy 2021, 234, 121275. [CrossRef]
10. Ghareshchopogh, F.S. Advances in Tree Seed Algorithm: A Comprehensive Survey. Arch. Comput. Methods Eng. 2022, 29, 3281–3304. [CrossRef]
11. Wang, H.R.; Li, R.; Xie, W.; Hao, J.J. The Research of Multi-Sensor Data Fusion for Medical Image Based on Weighted Least Square Method. Basic Clin. Pharmacol. Toxicol. 2016, 118, 30.
12. Paggi, H.; Soriano, J.; Lara, J.A. A Multi-Agent System for Minimizing Information Indeterminacy within Information Fusion Scenarios in Peer-to-Peer Networks with Limited Resources. Inf. Sci. 2018, 451–452, S1914198481. [CrossRef]
13. Barbouris, T.G.; Theocharis, J.B.; Alexiadis, M.C.; Dokopoulos, P.S. Long-Term Wind Speed and Power Forecasting Using Local Recurrent Neural Network Models. IEEE Trans. Energy Conver. 2006, 21, 273–284. [CrossRef]
14. Damousis, I.G.; Alexiadis, M.C.; Theocharis, J.B.; Dokopoulos, P.S. A fuzzy model for wind speed prediction and power generation in wind parks using spatial correlation. IEEE Trans. Energy Conver. 2004, 19, 352–361. [CrossRef]
15. Altan, A.; Karasu, S.; Zio, E. A new hybrid model for wind speed forecasting combining long short-term memory neural network, decomposition methods and grey wolf optimizer. Appl. Soft Comput. 2021, 100, 106996. [CrossRef]
16. Wu, C.; Wang, J.; Chen, X.; Du, P.; Yang, W. A novel hybrid system based on multi-objective optimization for wind speed forecasting. Renew. Energy 2020, 146, 149–165. [CrossRef]
17. Karasu, S.; Altan, A.; Sara, Z.; Hacioglu, R. Estimation of Fast Varied Wind Speed based on NARX Neural Network by using Curve Fitting. Int. J. Energy Appl. Technol. 2017, 4, 137–146.
18. Wang, Y.; Wang, J.; Li, Z.; Yang, H.; Li, H. Design of a combined system based on two-stage data preprocessing and multi-objective optimization for wind speed prediction. Energy 2021, 231, 121125. [CrossRef]
19. Ying, N.A.; Ni, L.B.; Jw, C. Ultra-short-term wind-speed bi-forecasting system via artificial intelligence and a double-forecasting scheme. Appl. Energy 2021, 301, 117452.
20. Lowndes, I.S.; Fogarty, T.; Yang, Z.Y. The application of genetic algorithms to optimise the performance of a mine ventilation network: The influence of coding method and population size. Soft Comput. 2005, 9, 493–506. [CrossRef]
21. Dembele, D.; Kastner, P. Fuzzy C-means method for clustering microarray data. Bioinformatics 2003, 19, 973–980. [CrossRef] [PubMed]
22. Chi, Y.; Jing, L. Reconstructing gene regulatory networks with a memetic-neural hybrid based on fuzzy cognitive maps. Nat. Comput. 2019, 18, 301–312. [CrossRef]
23. Zhou, J.; Chen, C.P.; Chen, L.; Li, H.X. A Collaborative Fuzzy Clustering Algorithm in Distributed Network Environments. *IEEE Trans. Fuzzy Syst.* 2013, 22, 1443–1456. [CrossRef]

24. Li, M.; Wang, Y.; Sun, Q.; Liu, Y. Research of ASW-FCM-Based Algorithm for Clustered Wind Turbine Group Equivalent Modeling. *J. Electr. Eng. Technol.* 2020, 15, 1555–1566. [CrossRef]

25. Ghafori, S.; Gharehchopogh, F.S. Advances in Spotted Hyena Optimizer: A Comprehensive Survey. *Arch. Comput. Methods Eng.* 2021, 29, 1569–1590. [CrossRef]

26. Li, M.; Wang, Y.; Sun, Q.; Liu, Y. Research of ASW-FCM-Based Algorithm for Clustered Wind Turbine Group Equivalent Modeling. *J. Electr. Eng. Technol.* 2020, 15, 1555–1566. [CrossRef]

27. Ghafori, S.; Gharehchopogh, F.S. Advances in Spotted Hyena Optimizer: A Comprehensive Survey. *Arch. Comput. Methods Eng.* 2021, 29, 1569–1590. [CrossRef]

28. Gharehchopogh, F.S. An Improved Tunicate Swarm Algorithm with Best-random Mutation Strategy for Global Optimization Problems. *J. Bionic Eng.* 2022, 19, 1177–1202. [CrossRef]

29. Gharehchopogh, F.S.; Shayanfar, H.; Gholizadeh, H. A comprehensive survey on symbiotic organisms search algorithms. *Artif. Intell. Rev.* 2020, 53, 2265–2312. [CrossRef]

30. Lan, T.; Liu, J.; Qin, H.; Xu, L.L. Time-domain global similarity method for automatic data cleaning for multi-channel measurement systems in magnetic confinement fusion devices. *Comput. Phys. Commun.* 2019, 234, 159–166. [CrossRef]

31. Li, C.; Lan, T.; Wang, Y.; Liu, J.; Xie, J.; Li, H.; Qin, H. An Automatic Data Cleaning Procedure for Electron Cyclotron Emission Imaging on EAST Tokamak Using Machine Learning Algorithm. *J. Instrum.* 2018, 13, P10029. [CrossRef]

32. Gharehchopogh, F.S.; Gholizadeh, H. A comprehensive survey: Whale Optimization Algorithm and its applications. *Swarm Evol. Comput.* 2019, 48, 1–24. [CrossRef]

33. Bose, A.; Mali, K. Type-reduced vague possibilistic fuzzy clustering for medical images. *Pattern Recognit.* 2021, 112, 107784. [CrossRef]

34. Durgarao, N.; Sudhavani, G. Diagnosing skin cancer via C-means segmentation with enhanced fuzzy optimization. *IET Image Process.* 2021, 15, 2266–2280. [CrossRef]

35. Fraiman, R.; Ghattas, B.; Svarc, M. Interpretable clustering using unsupervised binary trees. *Adv. Data Anal. Classif.* 2013, 7, 125–145. [CrossRef]