atmosphere

MDPI

Article

Research on Rapid Identification Technology of Sand and Dust Characteristic Monitoring Data Based on Optimized K-Means Clustering

Hao Zheng [1,2], Zhen Yang [2], Jianhua Yang [1,*], Linlin Zhang [2,3] and Yanan Tao [2]

[1] School of Automation, Northwestern Polytechnical University, Xi'an 710072, China
[2] Shaanxi Provincial Environmental Monitoring Center Station, Xi'an 710054, China
[3] China National Environmental Monitoring Centre, Beijing 100012, China
* Correspondence: 17391925643@163.com

Abstract: The criteria-based sand and dust weather determination method has the problem ofbeing a cumbersome and time-consuming process when processing a large amount of raw data, and cannot avoid the problems of repeatability and reproducibility. On the basis of statistical analysis of the air automatic monitoring data in the cities affected by sand and dust, this paper proposes a k-means optimization algorithm (MDPD-k-means) based on maximum density and percentage distance, which can quickly filter the characteristic data of sand and dust in a short time, and identify the days affected by sand and dust. This method effectively improves the data processing efficiency, solves the problems of poor reproducibility and large artificial error of traditional methods, and can support the business application of sand and dust data elimination. This paper uses the method to identify the sand and dust data of 10 cities in Shaanxi Province from 2016 to 2022, determines a total of 1107 sand and dust days, and points out that the number of days affected by sand and dust is increasing year by year. After excluding the effect of sand and dust, the urban PM_{10} concentration decreases by 18.42~1.41% respectively, which provides important data information for accurately evaluating the effectiveness of air pollution prevention and control.

Keywords: atmospheric environment monitoring network; sand and dust weather determination; k-means clustering optimization

Citation: Zheng, H.; Yang, Z.; Yang, J.; Zhang, L.; Tao, Y. Research on Rapid Identification Technology of Sand and Dust Characteristic Monitoring Data Based on Optimized K-Means Clustering. *Atmosphere* **2022**, *13*, 1720. https://doi.org/10.3390/atmos13101720

Academic Editors: Shan Huang and Wei Wei Hu

Received: 26 September 2022
Accepted: 13 October 2022
Published: 19 October 2022

1. Introduction

Sand and dust weather is a disastrous weather that affects the quality of the ecological environment. It is mainly produced by the soil wind erosion process in arid and semi-arid areas. Not only does it seriously threaten the surrounding ecological environment, human health, and industrial and agricultural production of sand source areas, but it also affects the regions on the transmission path to various degrees [1]. Sudden sand and dust weather leads to huge changes in air quality monitoring data. Indexes such as particle concentration, air-quality index (AQI), and comprehensive index far exceed the normal state during the sand and dust transmission process, and the air quality level often reaches heavy pollution or severe pollution. Different from the heavy pollution process with $PM_{2.5}$ as the primary pollutant in autumn and winter, when the sand and dust weather occurs, the mass concentrations of total suspended particulate (TSP) and PM_{10} in the ambient air increase rapidly in a short period of time. In severe cases, the PM_{10} mass concentration rapidly rises to above 1000 $\mu g/m^3$, and the AQI continues to be off the charts (AQI = 500), which seriously threatens human health [2]. Therefore, fast and accurate identification of sand and dust weather is of great significance for studying the law of sand and dust transmission, distinguishing heavy pollution processes from sand and dust transmission processes, and ensuring the stability of automatic monitoring network operation and the rationality of regional air quality evaluation.

Sand and dust sources in northern China are mainly distributed in the Hexi Corridor and Alxa region, the southern margin of the Southern Xinjiang Basin, and the central region of Inner Mongolia, whose central and western regions are one of the main sources of sand and dust in northwestern China [3,4]. In the absence of precipitation, the continuous increase in temperature in the sand source area leads to a decrease in the water content of the bare soil, which provides a material basis for the formation of sand and dust, while multiple strands of cold air move from north to south, resulting in strong winds and forming a transmission path for sand and dust diffusion [5–7]. With the development of atmospheric environment monitoring network, high temporal resolution monitoring data continue to accumulate, research on sand and dust characteristics continues to deepen, and sand and dust transmission paths and data characteristics become increasingly apparent. The area of effect and duration of sand and dust can be determined by identifying the start time, end time, peak concentration of particles, and physical properties of the particles [8,9].

For a single sand and dust transmission process, indexes including the depolarization ratio, extinction coefficient, and mixed layer height are directly observed by the space-borne lidar and ground-based lidar network, and these indexes are combined with the meteorological data and automatic particle monitoring data of the same period for analysis, in order to determine the area of effect and duration of sand and dust [10,11]. The extinction coefficient is the degree of attenuation of light by particles at a specific spatial coordinate point, which is usually positively related to the intensity of sand and dust. The depolarization ratio is a physical quantity that distinguishes spherical particles from non-spherical particles. The higher the proportion of non-spherical particles is, the more prominent the dust characteristics are [12]. When the extinction coefficient and depolarization ratio of near-surface observations increases rapidly, the dust transport height decreases or appears stratified. With the rapid settlement of coarse particles, the near-surface dust intensity reaches a peak and then gradually subsides [13]. When dust features are observed at high altitude, the dust process persists and continues to affect areas downstream of the transmission channel, and the backward trajectory model (HYSPLIT) is often used to analyze and verify the transmission path [14]. The quality of the data observed by lidar is greatly affected by natural conditions such as rain and snow weather, and the particle concentration cannot be directly observed. Therefore, it is usually used as an auxiliary support for ground-based particle concentration observation data in the process of sand and dust weather determination [15].

Aiming at the characteristics of sand and dust over a long time span, polar-orbiting and geostationary satellite remote sensing data, such as MODIS, Landsat, Himawari 8, FY-4, and CALIPSO are nested and superimposed, and the spectral characteristics of sand and dust particles in different regions are identified to determine sand sources, transmission paths, and affected areas, which are often used to identify the characteristics of sand and dust frequency changes in a large area over a long time span [16–19]. However, multispectral remote sensing methods are still restricted by objective factors such as cloud cover and land desertification, and the inversion results of dust intensity still need to be corrected by ground-based monitoring data.

With the rapid development of the automatic monitoring network of atmospheric environment, the monitoring data of particle concentration with high temporal resolution are accumulating continuously, which provides a lot of basic information for the judgment of sand and dust weather. The "Supplementary Regulations on the Evaluation of Urban Air Quality Affected by Sand and Dust Weather Processes" issued by the Ministry of Ecology and Environment of China in 2018 stipulates the method for determining sand and dust weather based on the criteria of PM_{10} and $PM_{2.5}$ hourly concentrations of the urban atmospheric environment monitoring network. The sand and dust process is determined by identifying when the sand and dust starts and ends. As a quantifiable identification method of sand and dust characteristics, the criteria method can determine the intensity and duration of sand and dust in affected cities. However, its determination process involves a large number of calculation, screening, judgment and audit links, in which the selection,

judgment and audit all rely on the experience of operators. Therefore, the method has poor reproducibility and repeatability in the face of massive historical monitoring data in a large number of cities.

Based on the criteria method, this paper proposes a fast identification method of sand and dust monitoring data based on optimized k-means clustering with the goal of supporting the business application of sand and dust identification. This method is used to study the sand and dust days in 10 cities in Shaanxi Province from 2016 to 2022, summarize the characteristics of air quality, and analyze the law of sand and dust transmission.

2. Materials and Methods

2.1. Criteria Method

The criteria method is mainly based on the hourly concentration changes of PM_{10} and $PM_{2.5}$ measured by the urban atmospheric environment monitoring network to determine the period affected by sand and dust. PM_{10} is the most important characteristic factor of sand and dust weather. Therefore, the hourly change characteristics of PM_{10} mass concentration are mainly considered when judging the impact of sand and dust. Sand and dust weather is usually accompanied by a sharp and rapid increase in PM_{10} mass concentration and a rapid decrease in the mass concentration ratio of $PM_{2.5}$ to PM_{10} [20,21]. In addition, taking into account the change characteristics of $PM_{2.5}$, the monitoring data of particulate matter in the period with obvious external sand and dust intrusion characteristics were analyzed, so as to identify the starting time and end time of the influence of sand and dust.

The starting time of sand and dust can be identified by criteria method. Either the time when the average urban PM_{10} hourly mass concentration is greater than or equal to twice the average PM_{10} mass concentration of the previous 6 h and greater than 150 $\mu g/m^3$ as the starting time of the sand and dust weather, or the time when the urban $PM_{2.5}$ to PM_{10} hourly mass concentration ratio is less than or equal to the previous 6 h 50% of the average hourly ratio, is taken as the starting time of the sand and dust weather.

The ending time of sand and dust can be identified by criteria method. Either the time when the hourly average mass concentration of PM_{10} in the city for the first time drops to a relative deviation of less than or equal to 10% from the average PM_{10} mass concentration in the previous 6 h before the sand and dust weather, or the moment when the hourly average mass concentration of PM_{10} in the city drops to less than 1.1 times of the average mass concentration of PM_{10} 6 h before the sand and dust weather for the first time, is taken as the ending time of the sand and dust weather.

The above judgment method is suitable for single-time sand and dust process identification in a single city with a small amount of data. When processing monitoring data of a long time, a large area, and multiple cities, the huge amount of data and the complex process will lead to a substantial increase in manual errors. In addition, the data characteristics of different regions are different, and the workload of data review and sand and dust weather determination is enormous, therefore it is difficult to obtain reliable and accurate statistical results in a short period of time.

2.2. Data Preprocessing

Reasonable data preprocessing can effectively improve the efficiency and accuracy of the clustering algorithm. When applying the distance-based clustering method, the mean and variance of the data set play a decisive role in the clustering results. Too many outliers make the clustering center shift, and some sand and dust data or outliers are far from the clustering center, which may lead to situations where the classification boundaries are blurred, and the critical point is difficult to accurately classify the classes. Therefore, before the cluster analysis, the original monitoring data need to be processed first, in order to further screen the target data and improve the data characteristics of the cluster center.

It can be seen from the determination process of the criteria method that the necessary conditions for the determination of sand and dust weather are the hourly concentration of PM_{10}, the concentration ratio of PM_{10} to $PM_{2.5}$, and the index of whether PM_{10} is the

primary pollutant of AQI. Only if the above three conditions are met at the same time, can the basic requirements for sand and dust data identification be met. Therefore, the hourly concentration of PM_{10}, the Individual Air Quality Index (IAQI) of PM_{10}, and the ratio of PM_{10} to $PM_{2.5}$ are analyzed as clustering elements.

1. Hourly concentration of PM_{10}. The characteristic pollutants of sand and dust weather are particulate matter (PM_{10}, $PM_{2.5}$), of which the short-term change of PM_{10} determines the strength of the sand and dust transmission process, therefore the hourly concentration of PM_{10} is an important condition for determining sand and dust weather. When the mass concentration of PM_{10} is greater than 150 $\mu g/m^3$, other data characteristics of sand and dust weather can be displayed.
2. IAQI of PM_{10}. When sand and dust weather occurs, PM_{10} is the only major pollutant, and the sub-index ($IAQI_{PM10}$) of PM_{10} is equal to AQI at this time. $IAQI_{PM10}$ is calculated based on the PM_{10} concentration in the original data, and whether it is the primary pollutant is identified by comparing with the AQI. Since the purpose of the experiment is to identify the sand and dust data for which PM_{10} is the only primary pollutant, in order to highlight the data characteristics, the index of whether PM_{10} is the primary pollutant is counted as 1 when PM_{10} is the primary pollutant, otherwise it is counted as 0, that is, when $IAQI_{PM10}$ = AQI, it is counted as 1; when $IAQI_{PM10}$ < AQI, it is counted as 0.
3. The concentration ratio of PM_{10} to $PM_{2.5}$. The concentration ratio of PM_{10} to $PM_{2.5}$ is another important factor in the determination of sand and dust weather. According to the PM_{10} and $PM_{2.5}$ air quality sub-indices and the concentration limits of corresponding pollutants given in the "Ambient Air Quality Index (AQI) Technical Regulations (Trial)" [22], it can be found that: When IAQI is 100, $C_{PM10}/C_{PM2.5}$ = 2; when IAQI is 150, $C_{PM10}/C_{PM2.5}$ is 2.17; when IAQI = 200, $C_{PM10}/C_{PM2.5}$ = 2.33; when IAQI is 300, $C_{PM10}/C_{PM2.5}$ = 1.68; when IAQI is 400, $C_{PM10}/C_{PM2.5}$ = 1.43; when IAQI is 500, $C_{PM10}/C_{PM2.5}$ = 1.2. When sand and dust weather occurs, air quality levels can range from mild to severe pollution, with AQI ranging from 100 to 500. If the primary pollutant is PM_{10}, the characteristic distribution of $C_{PM10}/C_{PM2.5}$ should be as shown in Figure 1.

In addition, the criteria method stipulates that "the hourly mass concentration ratio of $PM_{2.5}$ to PM_{10} is less than or equal to 50% of the average value of the ratio in the previous 6 h" [23], as shown in Equation (1):

$$\begin{cases} \frac{C_{PM10}}{C_{PM2.5}} = A_n \\ A_{n+7} \geq \overline{A_n} \times 2 \end{cases} \quad (1)$$

where C_{PM10} is the mass concentration of PM_{10}; $C_{PM2.5}$ is the mass concentration of $PM_{2.5}$; and A_n is the mass concentration ratio of PM_{10} to $PM_{2.5}$ at the nth hour.

Due to objective fact, $C_{PM10} \geq C_{PM2.5}$, therefore:

$$A_n \geq 1 \quad (2)$$

By substituting Equation (2) into (1), it can be deduced that:

$$A_{n+7} \geq 2 \quad (3)$$

Therefore, on the basis of satisfying the distribution law shown in Figure 1, the concentration ratio of PM_{10} to $PM_{2.5}$ should further satisfy the determination condition of $C_{PM10}/C_{PM2.5} \geq 2$.

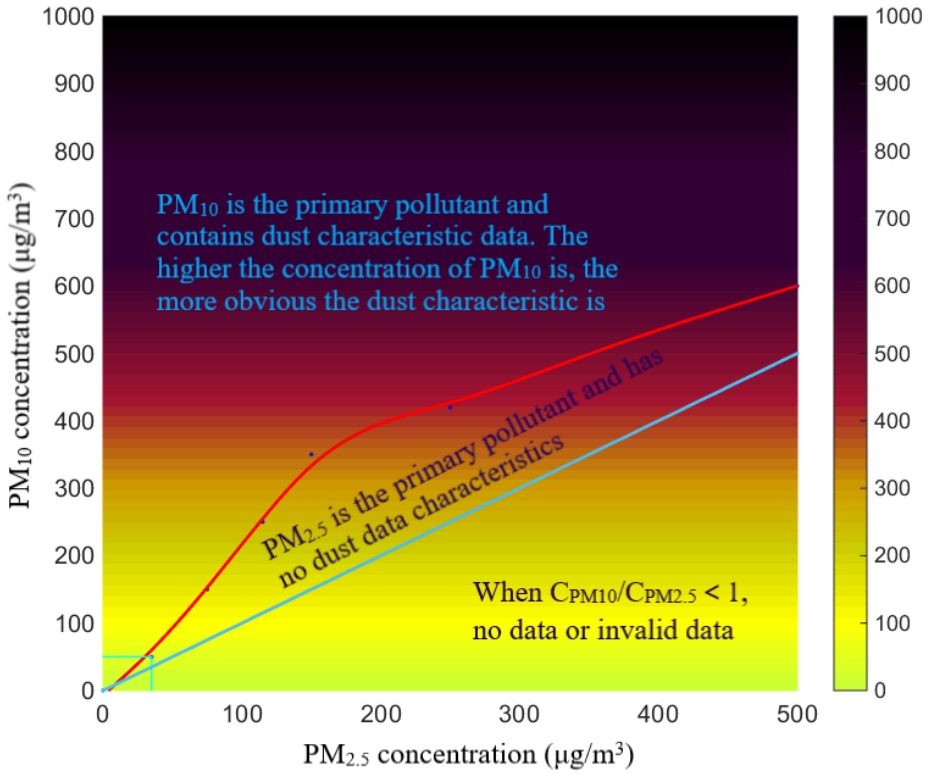

Figure 1. $C_{PM10}/C_{PM2.5}$ features distribution (The dark blue dots represent the concentration thresholds of PM_{10} and $PM_{2.5}$, corresponding to IAQI from 0 to 500; The red curve is fitted from dark blue points to distinguish the characteristics of PM_{10} or $PM_{2.5}$ as the primary pollutant; The light blue line indicates that PM_{10} concentration equals $PM_{2.5}$ concentration).

In summary, in order to further highlight the characteristics of sand and dust data, the experimental data used for cluster analysis are converted from the three indexes of PM_{10} concentration, $PM_{2.5}$ concentration, and AQI in the original data into PM_{10} concentration, the concentration ratio of PM_{10} to $PM_{2.5}$, and the index of whether PM_{10} is the primary pollutant (yes is 1, no is 0). The indexes settings are shown in Table 1.

Table 1. Index setting of sand dust data feature extraction.

Original Data	Converted Data	Basic Conditions for Determining Dust Data
C_{PM10}	C_{PM10}	$C_{PM10} > 150$ μg/m^3
$C_{PM2.5}$	$C_{PM10}/C_{PM2.5}$	$C_{PM10}/C_{PM2.5} \geq 2$
AQI	1 or 0	=1

Since the PM_{10} concentration and the concentration ratio of PM_{10} to $PM_{2.5}$ concentration are different dimensions, the data should be normalized first, transforming the variables into dimensionless numbers between [0, 1] by using the mapminmax function in MATLAB (The MathWork, Inc, Natick, MA, USA). For the case where the analysis process, especially the data processing process, involves data rounding, no data rounding should be performed during the calculation process, otherwise the original data information may be lost during denormalization.

2.3. K-Means Clustering

Based on the similarity of data features, cluster analysis divides similar objects in the data set into multiple categories, which is an exploratory classification process. Cluster analysis does not need to specify the classification criteria in advance, but can start from the data themselves, start unsupervised learning and perform clustering. Because the

characteristics of the same type of data are as similar as possible, and there are obvious differences in different types of data, the real distribution of the data can be analyzed in the end. In a data system with a stable operating system, the clusters of normal data are usually numerous and dense, while the clusters of abnormal data are small and sparse. Thus, the abnormal data can be preliminarily determined by the clustering method, and then other technical methods can be used to further analyze the data characteristics [24].

The k-means clustering algorithm is an iteratively solved partitioned clustering algorithm proposed by James MacQueen in 1967, which has the advantages of being simple, fast, and suitable for processing large-scale data. The basic idea is to randomly select k data objects from a data set containing a large number of data objects as the initial clustering centers and calculate the Euclidean geometric distance between each data object and the k clustering centers. All data are divided into the class represented by the cluster center closest to it, and the k cluster centers are updated according to the mean of the newly generated data objects in each category. If the change of the cluster center value in the adjacent iteration times exceeds the specified threshold, all data objects will be redivided according to the new cluster center; if the change of the cluster center value in the adjacent iteration times is less than the specified threshold, then the algorithm converges and the clustering result is output [25].

Since the initial cluster centers of k-means clustering are randomly selected, the final clustering results may vary. The k-means++ algorithm proposed by David Arthur in 2007 improves the selection of initial cluster centers based on k-means clustering [26]. Firstly, a sample from the data set is randomly selected as the cluster center C_n. According to the Euclidean geometric distance D_n between all samples and the cluster center C_n, the probability P_n that each sample is used as the next cluster center is calculated. Then a new initial cluster center C_{n+1} is randomly selected according to the probability. Finally, the above steps are repeated until k initial cluster centers appear, the clustering process of k-means is iterated to determine k final cluster centers and output the clustering result. The k-means++ algorithm is essentially the process of optimizing the initial clustering centers of the k-means algorithm, so that the k initial clustering centers can keep the maximum distance as much as possible, thereby improving the clustering accuracy and iterative efficiency, so as to obtain a relatively stable clustering result.

Since the k-means algorithm randomly selects the initial cluster centers, an unreasonable initial cluster center will affect the quality of the clustering results or the number of iterations of the algorithm. As an improved algorithm of k-means, k-means++ can separate the initial cluster centers as much as possible by screening the initial cluster centers, thereby enhancing the rationality of clustering and reducing the number of iterations. However, the first initial clustering center of the k-means++ algorithm is still randomly selected, and in the process of using the roulette method to select the initial clustering center of the target number, if there are many outliers, the clustering will still be affected. Therefore, the selection of initial clustering centers in the above two methods is random, and different clustering results may appear in the process of identifying sand and dust characteristic data, which is difficult to support the business application of the methods.

2.4. MDPD-k-Means Clustering

For the application scenario of feature recognition of sand and dust data, this paper proposes a K-means initial Clustering Center Optimization based on Maximum Density and Percentile Distance (MDPD-k-means). By dividing the coordinate grid in equal proportions and finding the center point of the maximum density grid, an initial cluster center is determined. By using the percentage distance instead of the roulette method, k initial clustering centers are selected, and the randomness of the algorithm is eliminated. When the number of clusters k is determined, the result will not change even if the clustering is performed multiple times, which can support the business application of the method.

First, the original data need to be dimensionally reduced. By replacing AQI to determination of whether PM_{10} is the primary pollutant, represented by 0 and 1, the data scatter

diagram for identifying sand and dust is transformed from one three-dimensional coordinate system to two two-dimensional coordinate systems. Then the two two-dimensional coordinate systems are combined to obtain a two-dimensional coordinate system, in which the information contained in each point is 0 or 1. Taking the monitoring results of Xi'an, Shaanxi Province from 1 February to 30 April 2018 as an example, the preprocessed data are shown in Figure 2.

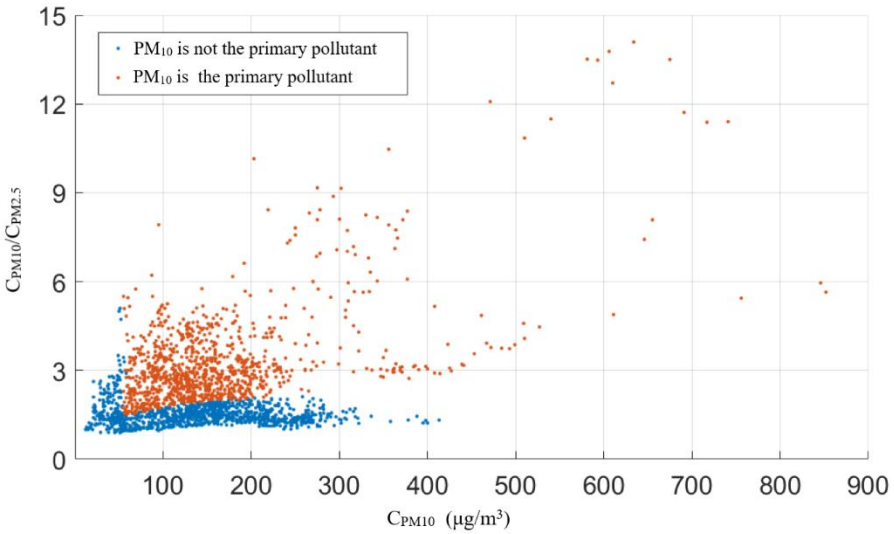

Figure 2. Scatter distribution of experimental data (Xi'an, 1 February to 30 April 2018).

After the data in Figure 2 are normalized to [0, 1] using MATLAB's mapminmax function, a grid of 0.1 × 0.1 is drawn to cover the entire coordinate system, and the number of data points in the 100 grids is calculated one by one and sorted. Since the characteristic data of sand and dust are few and sparse, and the characteristic data of non-sand dust are many and dense, the grid with the largest number of points must belong to the non-dust data set A_0. Therefore, the center coordinates C_0 of all points in A_0 are calculated and taken as the first initial cluster center.

When selecting the remaining initial cluster centers, the Euclidean distance from each point to each cluster center is calculated, the shortest distance D(x) is taken, and the percentile distance is sorted and obtained, until k initial cluster centers are selected. According to the sand emission frequency of sand source cities in Northwest China, taking the experience of identifying sand and dust data by the criteria method as a reference, the 95th percentile distance is used as the screening condition of the initial cluster center. Through the above improvements, the accuracy and convergence speed of clustering are improved, and the uncertainty is eliminated. The MDPD-k-means algorithm process can be summarized as follows:

Step1 Initializing: reading the dataset and the point information, and normalizing the dataset to make x, y ∈ [0, 1];

Step2 Dividing the density grid: dividing the two-dimensional coordinate system into 100 grids of 0.1 × 0.1, counting the number of scattered points in each grid, and sorting them according to the density from large to small;

Step3 Selecting the center point C1 of the grid with the largest number of scatter points, and calculating the distance D1(x) from all scatter points to C1;

Step4 Sorting D1(x) from small to large, selecting a point C2 at the 95th percentile distance, calculating the distances from all scattered points to C1 and C2, and taking the minimum distance D12(x) from each point to C1 and C2;

Step5 Sorting D12(x) from small to large, selecting a point C3 at the 95th percentile distance to calculate the distances from all scatter points to C1, C2, and C3, and taking the minimum distance D123(x) from each point to C1, C2, and C3;

Step6 Repeating step 5 until k center points appear, and outputting k center point coordinates;

Step7 Performing k-means clustering with k center points as the initial cluster centers, and outputting the number of iterations, the final cluster center, the clustering results and the number of cases;

Step8 After de-normalization, carrying out feature determination based on case information.

3. Results and Discussion

3.1. Results of Clustering

Taking the ambient air quality monitoring data of Xi'an City, Shaanxi Province from 1 February to 30 April 2018, as the research object, including the hourly mean concentrations of PM_{10} and $PM_{2.5}$, and the city's hourly AQI, a total of 2136 groups of raw data were extracted from the Shaanxi Provincial Ambient Air Quality Monitoring Network Management Platform. After the data were preprocessed, MATLAB was used to run the MDPD-k-means clustering program, the number of clusters was determined and set to 7 according to the elbow rule, and the 95th percentile distance was taken. The clustering results converged after 22 iterations, and by the denormalization of the clustering results, seven cluster centers were obtained as shown in Table 2, and the clustering results were shown in Figure 3.

Table 2. Clustering centers of MDPD-k-means algorithm.

MDPD-k-Means Clustering Results							
Categories	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7
C_{PM10}	212.55	84.44	148.37	290.77	120.01	390.63	654.35
$C_{PM10}/C_{PM2.5}$	1.47	1.53	3.81	7.43	2.45	3.50	10.74
1 or 0	0	0	1	1	1	1	1
Case number	494	592	357	47	577	52	17

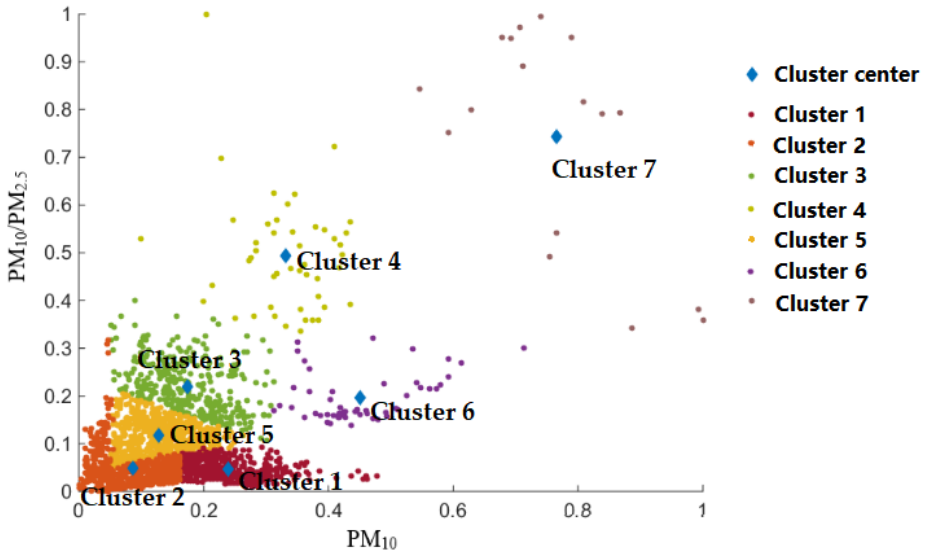

Figure 3. Clustering sample distribution of MDPD-K-means algorithm (Coordinate system based on the normalized scale).

Using the three basic characteristics of the sand and dust data in Table 1 to determine whether the cluster centers in Table 2 conform to the characteristics of the dust data one by one, it could be found that Cluster 4, Cluster 6, and Cluster 7 generated by the MDPD-k-means algorithm conform to the three basic characteristics at the same time, while at least one of the indexes in Cluster 1, Cluster 2, Cluster 3, and Cluster 5 did not meet the

characteristics of sand and dust, and could not be identified as sand and dust data. The identification results were shown in Table 3 and Figure 4 for details.

Table 3. The clustering results and dust data judgment of MDPD-k-means algorithm.

Clusters	Case Count	Percentage of Cases	Dust Characteristics of Cluster Centers			Judgment
			C_{PM10}	$C_{PM10}/C_{PM2.5}$	1 or 0	
Cluster 4	47	2.2%	290.77	7.43	1	Dust characteristics
Cluster 6	52	2.4%	390.63	3.50	1	
Cluster 7	17	0.8%	654.35	10.74	1	
Cluster 1	494	23.1%	212.55	1.47	0	Non dust characteristics
Cluster 2	592	27.7%	84.44	1.53	0	
Cluster 3	357	16.7%	148.37	3.81	1	
Cluster 5	577	27.0%	120.01	2.45	1	

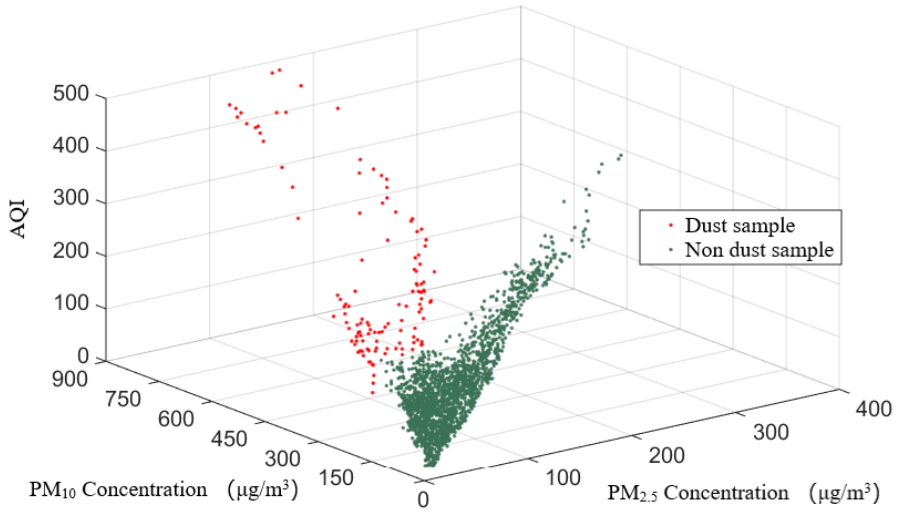

Figure 4. Dust samples identified by MDPD-k-means algorithm.

3.2. Accuracy Analysis

Taking the identification results of the traditional criteria method as the real classification, the confusion matrix was used to evaluate the accuracy of the MDPD-k-means clustering results, and compared with the k-means, k-means++ and DBSCAN clustering results. Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is a density-based unsupervised machine learning clustering algorithm, which is suitable for detecting outliers in samples. Firstly, the similarity between the recognition results of the four groups of clustering algorithms for sand and dust data and the 118 groups of data recognized by the criteria method was compared. Since the purpose of the experiment was to identify the characteristic data of sand and dust, the identification results of the above four algorithms were divided into two categories according to the sand-dust data and the non-sand-dust data. Next, the intersect function of MATLAB was used to compare the cross relationship between the sand and dust characteristic data identified by the above four clustering algorithms and the sand and dust characteristic data identified by the criteria method, as shown in Table 4.

As could be seen from Table 4, among the 116 sets of dust data identified by the MDPD-k-means algorithm, 109 sets of data were consistent with the judgment results of the criteria method, and 7 sets of data were identified incorrectly. The seven sets of identification error data all appeared in the initial stage of the increase of PM_{10} concentration and the decrease of $PM_{2.5}$ concentration, that is, the continuous occurrence of haze days and dusty days, and the rapid transition stage of $PM_{2.5}$ pollution to PM_{10} pollution. A total of nine sets of dust characteristic data were not effectively identified. The main reason was that these data were located at the end of the sand and dust transmission process, and the pollutant

concentration began to decline rapidly and tended to normal levels, but the data conformed to the characteristics of sand and dust, thus the clustering algorithm could not accurately classify categories. Although the MDPD-k-means algorithm did not identify all the sand and dust data identified by the criteria method, its recognition rate of 92.37% was still higher than that of k-means, k-means++, DBSCAN and other algorithms, which can better support the sand and dust characteristic data quick identification.

Table 4. Comparison between the recognition results of 4 algorithms and criterion method.

Methods	Dust Data (Group)	Non Dust Data (Group)	Number of Samples Intersecting with Criterion Data (Group)		
			Overall Data Intersection	Dust Data Intersection	Non Dust Data Intersection
k-means	122	2014	2100	101	1999
k-means++	114	2022	2103	101	2002
DBSCAN	102	2034	2070	75	1995
MDPD-k-means	116	2020	2120	109	2011

Using confusion matrix to evaluate the accuracy of the four clustering methods for sand and dust data identification, the confusion matrix of the identification results of the four clustering methods and the true value (criteria method results) were drawn respectively, as shown in Figure 5.

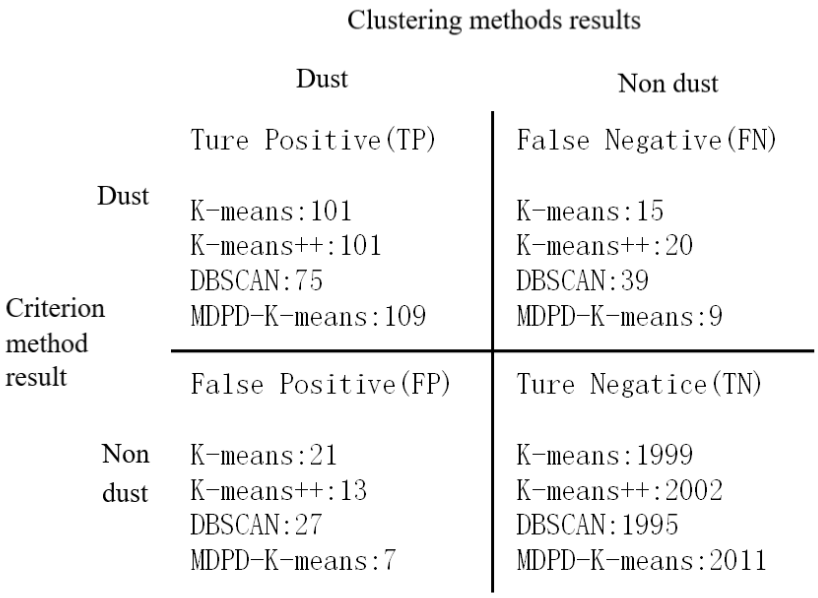

Figure 5. Confusion matrix of case number by using clustering method and criterion method.

According to the confusion matrix in Figure 5, the ratio of the number of correctly identified sand and dust samples to the total number of actual sand and dust samples (*TPR*), the ratio of the number of incorrectly identified sand and dust samples to the total number of actual sand and dust samples (*FNR*), the ratio of the number of incorrectly identified non-sand-dust samples to the total number of non-sand-dust samples (*FPR*), and the ratio of the number of correctly identified non-sand-dust samples to the total number of actual non-sand-dust samples (*TNR*) can all be calculated. The formula is shown as follows:

$$\begin{array}{l} TPR = \frac{TP}{TP+FN} \\ FNR = \frac{FN}{TP+FN} \\ FPR = \frac{FP}{FP+TN} \\ TNR = \frac{TN}{FP+TN} \end{array} \quad (4)$$

Based on *TP*, *FN*, *FP*, *TN*, the overall clustering recognition Accuracy and sand and dust recognition Precision can be calculated as follows:

$$\begin{array}{c} Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \\ Precision = \frac{TP}{TP+FP} \end{array} \quad (5)$$

*F*1_*score* is calculated according to Equations (4) and (5):

$$F1\_score = 2 \times \frac{Precision \times TPR}{Precision + TPR} \quad (6)$$

*F*1_*score* is a standard of measurement of classification problems, and its value ranges from 0 to 1, where 0 represents the worst output of the model, and 1 represents the best.

The *TPR*, *FNR*, *FPR*, *TNR*, accuracy, precision and *F*1_*score* calculation results of the four clustering methods were shown in Table 5.

Table 5. Comparison of evaluation indexes of 4 clustering methods.

Methods	Evaluation Indexes						
	TPR	***FNR***	***FPR***	***TNR***	**Accuracy**	**Precision**	***F1_Score***
K-means	87.07%	12.93%	1.04%	98.96%	98.31%	82.79%	0.85
K-means++	83.47%	16.53%	0.65%	99.35%	98.46%	88.60%	0.86
DBSCAN	65.79%	34.21%	1.34%	98.66%	96.91%	73.53%	0.69
MDPD-k-means	92.37%	7.63%	0.35%	99.65%	99.25%	93.97%	0.93

It could be seen from Table 5 that k-means and its optimization method were suitable for the rapid identification of sand and dust feature data, among which the MDPD-k-means algorithm was better than other clustering algorithms in terms of recognition accuracy, precision and *F*1_*score* for sand and dust feature data, with its determined sand and dust characteristic data closest to that of the criteria method. For the hourly data of sand and dust identified by the MDPD-k-means algorithm, according to the principle that if the impact time $\geq$ 1 h, that day is a day affected by sand and dust, 9 days affected by sand and dust were obtained, which was exactly the same as the date identified by the criteria method. Therefore, the MDPD-k-means algorithm has high accuracy and efficiency in identifying the sand and dust features of hourly monitoring data and can support the business application of sand and dust feature recognition.

3.3. Applicability Analysis

The criteria method is supported by the Ministry of Ecology and Environment's "Supplementary Regulations on the Evaluation of Urban Air Quality Affected by Sand-dust Weather Processes" and is being widely used in practical work. It is the process of calculating the characteristic data and then filtering the combination, and the original data are dimensionally reduced by calculation and screening to meet the requirements of data feature identification. However, the criteria determination process involves a large number of calculation, screening, judgment and review processes, among which screening, judgment and review are easily affected by the subjective influence of operators, especially when processing monitoring data sets under critical conditions. There are differences in the results judged by different operators based on their different experience. In the face of massive historical monitoring data in a large number of cities, this method is time-consuming and labor-intensive, requires high operator experience, and has poor reproducibility of the determination process.

The technical idea of clustering to identify sand and dust characteristic monitoring data is different from that of criteria method. There is no manual calculation and screening process at the data level. Clustering all original monitoring data according to data features through unsupervised machine learning is a process of data classification and re-identification of features, and the original data are dimensionally reduced by data classification to meet the requirements of data feature judgment. In this process, the classification is completed by the clustering algorithm of the computer software, which can process a large amount of monitoring data at the same time according to the unified algorithm, so the clustering result will not lose the original data information, the process reproducibility is strong, and there will not be discrepancies in judgment due to the inexperience of the staff in the data calculation and classification process. In the partition-based clustering method represented by the k-means algorithm, there is no uniform standard for the number of clusters k, and it is easy to miss the data features with relatively low mass concentrations of PM_{10} and $PM_{2.5}$, for example the end time of sand and dust weather is determined earlier than that determined by the criteria method. In addition, whether it is the k-means or k-means++ algorithm, there is a certain degree of randomness in the selection of the initial cluster center. When the preset indicators are the same, the results obtained by running the algorithm multiple times may be inconsistent and cannot support business application.

The MDPD-k-means algorithm eliminates the randomness in the process of selecting the initial cluster center by the k-means or k-means++ algorithm, ensures that the cluster center can be classified into the sand and dust data through the density grid and percentile distance, and improves the efficiency of feature extraction and reduces the number of iterations, which can support the business application of sand and dust identification. Compared with the criteria method, the use of MDPD-k-means clustering algorithm can minimize the workload of manual judgment and avoid manual errors and systematic errors to the greatest extent, while the determination process has better reproducibility and repeatability, therefore can be used for simultaneous identification of multi-region and large-scale hourly monitoring data. However, the k value still affects the clustering performance. When a large number of samples are processed, the amount of computation will increase rapidly as the value of k increases. The recommended sample size is continuous 720 h (1 month) to 8760 h (1 year), the number of k values is 5–10. The model operates quickly and the results are accurate.

4. Identification of Historical Sand and Dust Data

4.1. Sand and Dust Weather Determination Results

A total of 10 cities in Shaanxi Province were used as research objects, and the data source was the real-time PM_{10} and $PM_{2.5}$ hourly concentration status of 50 air automatic stations in 10 cities from 1 January 2016 to 31 May 2022 obtained by the "Shaanxi Provincial Ambient Air Quality Monitoring Network Management Platform". According to the distribution of stations in each city, the city hourly mean value was calculated, and the hourly AQI was counted, obtaining a total of 1.703 million pieces of hourly PM_{10}, $PM_{2.5}$, and AQI data of 10 cities from 2016 to 2022.

The 1.703 million pieces of data were divided into 10 groups according to different cities, and the MDPD-k-means clustering method was used to identify the hourly data of 10 cities in Shaanxi Province from 2016 to 2022 that conformed to the characteristics of sand and dust. According to the clustering results, the monitoring data conformed to the characteristics of sand and dust in each city was screened out, and the sand-dust-affected days of 10 cities from 2016 to 2022 were determined daily according to the corresponding time. Since the data were identified based on the characteristics of hourly values, according to the principle that if the influence time $\geq$1 h, that day is a day affected by sand and dust, the final determination of the number of days affected by sand and dust in 10 cities in Shaanxi Province from 2016 to 2022 was shown in Table 6.

Table 6. Dust days of 10 cities in Shaanxi Province from 2016 to 2022 by using MDPD-k-means method.

City	Dust Affected Days							Total
	January to December 2016	January to December 2017	January to December 2018	January to December 2019	January to December 2020	January to December 2021	January to May 2022	
Xi'an	8	7	18	15	15	41	31	135
Baoji	5	4	16	11	11	33	21	101
Xianyang	9	5	22	17	17	41	32	143
Tongchuan	9	9	17	8	16	32	26	117
Weinan	10	9	24	19	17	39	33	151
Yanan	9	6	23	14	22	40	36	150
Yulin	10	8	26	16	29	52	35	176
Hanzhong	5	1	4	3	5	7	7	32
Ankang	2	0	4	2	2	7	4	21
Shangluo	6	7	12	6	12	22	16	81
Average	7.3	5.6	16.6	11.1	14.6	31.4	24.1	110.7

4.2. Characteristics of Data Changes

After excluding the impact of sand and dust, the annual average PM_{10} concentration in 10 cities has changed to varying degrees. It can be seen from Figure 6 that from 2016 to 2022, northern Shaanxi is heavily affected by sand and dust. Among them, Yulin is located in the transition zone between the Loess Plateau and the Inner Mongolia Plateau, and is close to the Mu Us Sand source, therefore is most seriously affected by sand and dust, with its PM_{10} concentration dropping by 18.42% after excluding the impact of sand and dust. Yan'an is severely affected, with its PM_{10} concentration dropping by 12.99%. Hanzhong and Ankang in southern Shaanxi are less affected by sand and dust, and their PM_{10} concentrations decrease by 1.41–1.69% after excluding the impact of sand and dust. Shangluo City is at the end of the sand and dust transmission path. Due to the low background value of its urban environment, it is prominently affected by sand and dust, and its PM_{10} concentration drops by 8.62% after excluding the impact of sand and dust. Cities in the Guanzhong area all belong to the area affected by sand and dust transmission. Due to the combined effect of sand and dust transmission and local fugitive dust, its PM_{10} concentration decreases by 5.68–7.14% after excluding the impact of sand and dust.

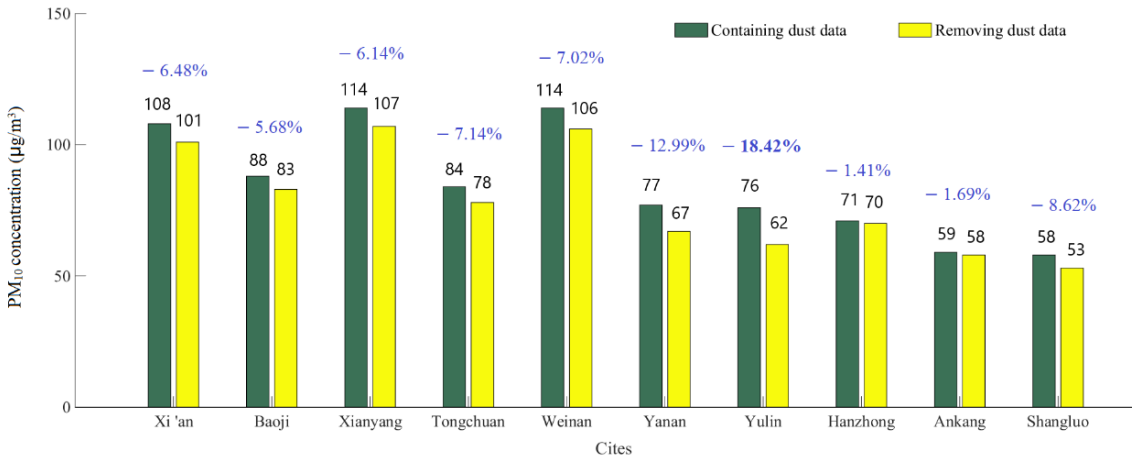

Figure 6. Comparison of PM_{10} average concentrations containing and removing dust data in 10 cities from 2016 to 2022.

ArcGIS was used to draw the distribution map of sand and dust days in Shaanxi Province from 2016 to 2021. As shown in Figure 7, from 2016 to 2021, 10 cities in Shaanxi Province experienced sand and dust weather to varying degrees, and the number of days with sand and dust showed an overall upward trend, while the increase was obvious in 2021. The number of times of cities affected by sand and dust gradually increased from south to north. Yulin and Yan'an were frequently affected by sand and dust, and the

number of times of impact was increasing year by year. Cities in the Guanzhong region were seriously affected by sand and dust from 2018 to 2020, and the number of sand and dust days was basically the same every year, but it increased significantly in 2021. The southern Shaanxi region was less affected by sand and dust. However, located at the end of the sand and dust transmission path, the number of affected times of Shangluo was higher than that of the other two cities in southern Shaanxi.

From the analysis of topography and transmission path, the sand and dust in the northwest originate from the Hexi Corridor in Gansu and enter the Guanzhong Plain from west to east through Tianshui–Baoji. At the same time, affected by the return of sand and dust in the east, sand and dust remain in Guanzhong and accumulate, causing secondary pollution to the city of Guanzhong, showing the characteristics of intermittent occurrence and decreasing intensity of sand and dust for several consecutive days. Originating from Inner Mongolia and Ningxia, the northern sand and dust travel south through Yulin and Yan'an, enter the Guanzhong Plain from north to south, resulting in a rapid increase in PM_{10} concentrations in cities along the way, and cross the Qinling Mountains to affect the southern Shaanxi area, mainly Shangluo, showing large-scale, high-intensity sand and dust transport characteristics. The transmission of other northern sand and dust starts from Mongolia and travels south through the Beijing–Tianjin–Hebei region. At the end of the sand and dust transmission process, it usually enters the Guanzhong Plain from Shanxi, which has a certain impact on Xi'an, Xianyang, and Weinan. Through years of sand control and soil erosion control in Shaanxi Province, the impact of local sand and dust has been basically eliminated, and the PM_{10} concentration has dropped significantly. However, many cities are located in the sand and dust transmission channels, and the sand and dust transmission process leads to a short-term rapid increase in PM_{10} concentration. Although the concentration of PM_{10} can be reduced through measures such as regional air pollution prevention and control, it is difficult to effectively reduce the number of sandy and dusty days.

It could be seen from the frequency and distribution of sand and dust weather from 2016 to 2022 shown in Figure 8 that Shaanxi Province was mainly affected by sand and dust from March to May, and secondarily affected by sand and dust from November to December. The transmission of sand and dust led to a rapid increase in the concentration of PM_{10}, with more pollution days and heavier pollution level. March was the month with the highest frequency of sand and dust occurrences, accounting for 29.7% of all sand and dust days in Shaanxi Province from 2016 to 2022. Sorting the daily average PM_{10} concentration of each sand and dust process from large to small, the highest value of PM_{10} daily average concentration appeared in Yulin City, which was 3673 $\mu g/m^3$ on 15 March 2021, followed by 2980 $\mu g/m^3$ of Yan'an City on 16 March. In addition, in 2021, Yulin and Yan'an had four sand and dust days with an average daily PM_{10} concentration of over 1000 $\mu g/m^3$, which seriously threatened human health. In Tongchuan, Baoji, Xianyang, Weinan, Xi'an and Shangluo, the average daily concentration of PM_{10} exceeded 600 $\mu g/m^3$ dust for many times, which seriously affected the ambient air quality. The one with the widest impact was the sand and dust transmission process from 12 to 14 May 2019, which affected all cities in Shaanxi Province, and the average daily PM_{10} concentrations in Shangluo on May 12 and 13 were 626 $\mu g/m^3$ and 499 $\mu g/m^3$. This sand and dust process caused the average annual concentration of PM_{10} in Shangluo to increase by 2.7 $\mu g/m^3$ in 2019, while the sand dust in the whole year of 2019 caused the average annual concentration of PM_{10} in Shangluo to increase by 4.0 $\mu g/m^3$ to 58 $\mu g/m^3$, causing the excess of the standard (>70 $\mu g/m^3$) of PM_{10} concentration in the three cities of Yulin, Yan'an and Hanzhong, which was not conducive to air quality evaluation and national ranking. Moreover, the transmission of sand and dust greatly increased the air quality level. From 2016 to 2022, there were 142 days of severe and above pollution caused by sand and dust, accounting for 12.7% of the total number of days with sand and dust, which was not conducive to the reduction of heavily polluted weather. When there is no effective method to control the sand emission conditions in the northern sand source areas, sand prevention and dust suppression measures can be taken to reduce the superimposed pollution of sand and dust transmission and local particle sources. Based

on the overall changes in air quality in 10 cities in Shaanxi Province from 2016 to 2022, it can be found that although the ambient air quality shows an overall improvement trend, with the increase in the number of sand and dust occurrences year by year, the proportion of sand and dust transmission on air quality will further increase, thereby reversing the situation of air quality improvement.

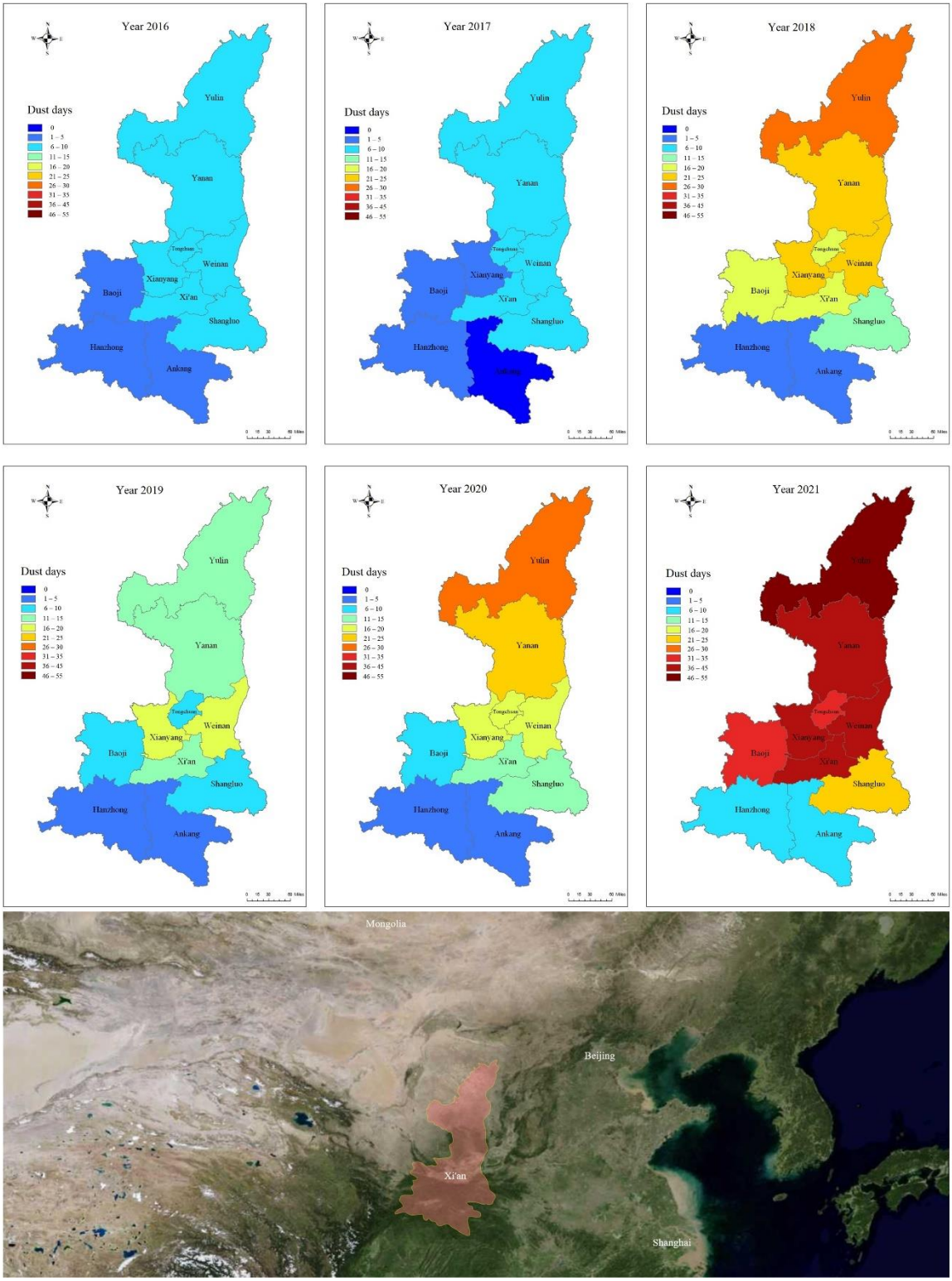

Figure 7. Sand affected days in Shaanxi Province from 2016 to 2021.

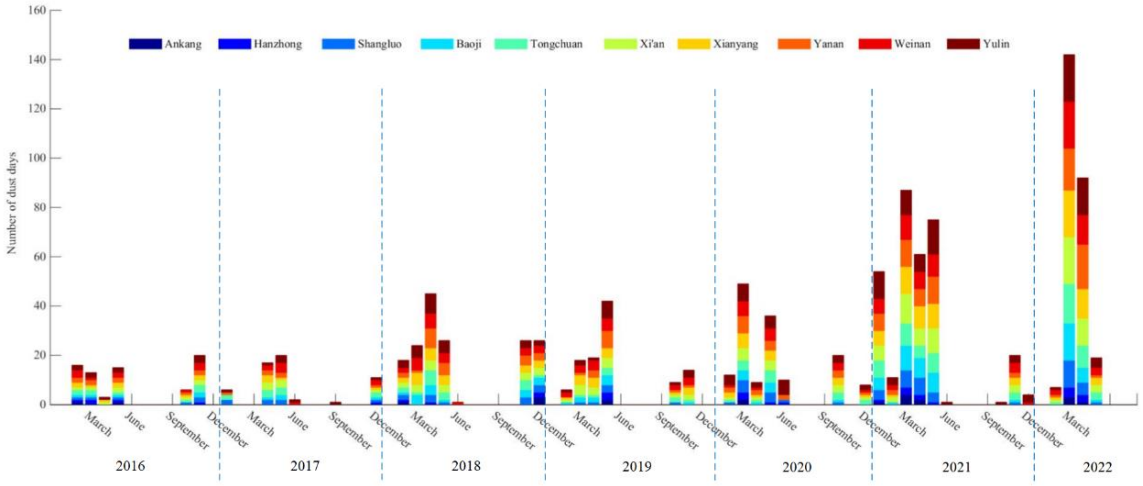

Figure 8. Frequency and distribution of dust weather from 2016 to 2022.

5. Conclusions

Aiming at the problems of cumbersome and time-consuming process, unavoidable repeatability and reproducibility errors when processing massive raw data of atmospheric environment monitoring network based on traditional sand and dust weather determination methods, this paper discusses the feasibility of using clustering algorithm to identify sand and dust data, optimizes the k-means clustering algorithm, and proposes a MDPD-k-means algorithm based on the maximum density and percentage distance, forming a relatively complete sand and dust data identification process which can quickly process a large amount of raw data in a short time. The determination efficiency of the proposed method for sand and dust data is much higher than that of the criteria method, which effectively solves the problems of cumbersome calculation process, poor process reproducibility, and large manual error when faced with a large amount of original data, and the recognition accuracy is also higher than other clustering methods. The proposed method is suitable for the business application of sand and dust data elimination, and has strong supporting significance for the research and judgment of regional atmospheric pollution situation.

In addition, the MDPD-k-means algorithm is used to identify the characteristics of sand and dust on the hourly data of the atmospheric environment monitoring network in 10 cities in Shaanxi Province from 2016 to 2022. According to the principle that if the impact time is more than 1 h, the day is a day affected by sand and dust, a total of 1107 sand and dust days are identified, including 142 days with severe and above pollution, and the daily average concentration of PM_{10} exceeded 600 $\mu g/m^3$ sand for many times, which seriously affects the ambient air quality and threatens human health. It can be found from the changes in the days affected by sand and dust that the number of sand and dust weather occurrences in 10 cities in Shaanxi Province shows an overall upward trend from 2016 to 2022, and a larger increase from 2021 to 2022. As the concentration of air pollutants continues to decline, the proportion of sand and dust transport on air quality will further increase, which will have a serious impact on the improvement of air quality and is not conducive to reducing heavily polluted weather. After eliminating the impact of sand and dust, the PM_{10} concentrations in 10 cities in Shaanxi Province decreased by 18.42%~1.41% respectively, providing important data information for accurate assessment of the effectiveness of air pollution prevention and control and assessment of ambient air quality.

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

Funding: This research was supported by the Natural Science Basis Research Plan in Shaanxi Province of China (2021JQ-963).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

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

References

1. Xu, J. Sand-dust storms in and around the Ordos Plateau of China as influenced by land use change and desertification. *CATENA* **2006**, *65*, 279–284. [CrossRef]
2. An, X.; Hou, Q.; Li, N.; Zhai, S. Assessment of human exposure level to PM_{10} in China. *Atmos. Environ.* **2013**, *70*, 376–386. [CrossRef]
3. Zhang, K.; Chai, F.; Zhang, R.; Xue, Z. Source, route and effect of Asian sand dust on environment and the oceans. *Particuology* **2010**, *8*, 319–324. [CrossRef]
4. Wang, X.; Dong, Z.; Zhang, J.; Liu, L. Modern dust storms in China: An overview. *J. Arid. Environ.* **2004**, *58*, 559–574. [CrossRef]
5. Yang, Y.Q.; Hou, Q.; Zhou, C.H.; Liu, H.L.; Wang, Y.Q.; Niu, T. Sand/dust storm processes in Northeast Asia and associated large-scale circulations. *Atmos. Chem. Phys.* **2008**, *8*, 25–33. [CrossRef]
6. Kimura, R. Factors contributing to dust storms in source regions producing the yellow-sand phenomena observed in Japan from 1993 to 2002. *J. Arid. Environ.* **2012**, *80*, 40–44. [CrossRef]
7. Yang, Y.; Qu, Z.; Shi, P.; Liu, L.; Zhang, G.; Tang, Y.; Hu, X.; Lv, Y.; Xiong, Y.; Wang, J.; et al. Wind regime and sand transport in the corridor between the Badain Jaran and Tengger deserts, central Alxa Plateau, China. *Aeolian Res.* **2014**, *12*, 143–156. [CrossRef]
8. Mao, J.; Sheng, H.; Zhao, H.; Zhou, C. Observation Study on the Size Distribution of Sand Dust Aerosol Particles over Yinchuan, China. *Adv. Meteorol.* **2014**, *6*, 1–7. [CrossRef]
9. Shimizu, A.; Nishizawa, T.; Jin, Y.; Kim, S.-W.; Wang, Z.; Batdorj, D.; Sugimoto, N. Evolution of a lidar network for tropospheric aerosol detection in East Asia. *Opt. Eng.* **2017**, *56*, 031219. [CrossRef]
10. Mona, L.; Liu, Z.; Muller, D.; Omar, A.; Papayannis, A.; Pappalardo, G.; Sugimoto, N.; Vaughan, M. Lidar Measurements for Desert Dust Characterization: An Overview. *Adv. Meteorol.* **2012**, *7*, 1449–1458. [CrossRef]
11. Todd, M.C.; Cavazos-Guerra, C. Dust aerosol emission over the Sahara during summertime from Cloud-Aerosol Lidar with Orthogonal Polarization (CALIOP) observations. *Atmos. Environ.* **2016**, *128*, 147–157. [CrossRef]
12. Luo, T.; Wang, Z.; Ferrare, R.A.; Hostetler, C.A.; Yuan, R.; Zhang, D. Vertically resolved separation of dust and other aerosol types by a new lidar depolarization method. *Opt. Express* **2015**, *23*, 14095. [CrossRef] [PubMed]
13. Han, Y.; Wu, Y.; Wang, T.; Xie, C.; Zhao, K.; Zhuang, B.; Li, S. Characterizing a persistent Asian dust transport event: Optical properties and impact on air quality through the ground-based and satellite measurements over Nanjing, China. *Atmos. Environ.* **2015**, *115*, 304–316. [CrossRef]
14. Zhang, J.; Chen, Z.; Lu, Y.; Gui, H.; Liu, J.; Liu, W.; Wang, J.; Yu, T.; Cheng, Y.; Chen, Y.; et al. Characteristics of aerosol size distribution and vertical backscattering coefficient profile during 2014 APEC in Beijing. *Atmos. Environ.* **2017**, *148*, 30–41. [CrossRef]
15. Ceolato, R.; Berg, M.J. Aerosol light extinction and backscattering: A review with a lidar perspective. *J. Quant. Spectrosc. Radiat. Transf.* **2021**, *262*, 107492. [CrossRef]
16. Rayegani, B.; Barati, S.; Goshtasb, H.; Gachpaz, S.; Ramezani, J.; Sarkheil, H. Sand and dust storm sources identification: A remote sensing approach. *Ecol. Indic.* **2020**, *112*, 106099. [CrossRef]
17. Bao, Y.; Zhu, L.; Guan, Q.; Guan, Y.; Lu, Q.; Petropoulos, G.; Che, H.; Ali, G.; Dong, Y.; Tang, Z.; et al. Assessing the impact of Chinese FY-3/MERSI AOD data assimilation on air quality forecasts: Sand dust events in northeast China. *Atmos. Environ.* **2019**, *205*, 78–89. [CrossRef]
18. Guo, J.; Niu, T.; Wang, F.; Deng, M.; Wang, Y. Integration of multi-source measurements to monitor sand-dust storms over North China: A case study. *Acta Meteorol. Sin.* **2013**, *27*, 566–576. [CrossRef]
19. Zhao, S.; Yin, D.; Qu, J. Identifying sources of dust based on CALIPSO, MODIS satellite data and backward trajectory model. *Atmos. Pollut. Res.* **2015**, *6*, 36–44. [CrossRef]

20. Sugimoto, N.; Shimizu, A.; Matsui, I.; Nishikawa, M. A method for estimating the fraction of mineral dust in particulate matter using $PM_{2.5}$-to-PM_{10} ratios. *Particuology* **2016**, *28*, 114–120. [CrossRef]
21. Li, J.; Wang, S.; Chu, J.; Wang, J.; Li, X.; Yue, M.; Shang, K. Characteristics of air pollution events over Hotan Prefecture at the southwestern edge of Taklimakan Desert, China. *J. Arid. Land* **2018**, *10*, 686–700. [CrossRef]
22. *HJ633-2012*; Ambient Air Quality Index (AQI) Technical Regulations (Trial). National Environmental Protection Standards of the People's Republic of China: Beijing, China, 2012. Available online: https://www.mee.gov.cn/ywgz/fgbz/bz/bzwb/jcffbz/201203/t20120302_224166.shtml (accessed on 1 January 2016).
23. Ministry of Ecology and Environment of the People's Republic of China. Supplementary Regulations on the Evaluation of Urban Air Quality Affected by Sand-dust Weather Processes. 2018. Available online: http://www.mee.gov.cn/gkml/hbb/bgt/201701/t20170106_394054.htm (accessed on 4 January 2017).
24. Nagpal, A.; Jatain, A.; Gaur, D. Review based on data clustering algorithms. In Proceedings of the 2013 IEEE Conference on Information and Communication Technologies, Thuckalay, India, 11–12 April 2013; pp. 298–303. [CrossRef]
25. Taher, N.; Babak, A. An efficient hybrid approach based on PSO, ACO and k-means for cluster analysis. *Appl. Soft Comput.* **2010**, *10*, 183–197. [CrossRef]
26. Arthur, D.; Vassilvitskii, S. K-Means++: The Advantages of Careful Seeding. In Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA, New Orleans, LA, USA, 7–9 January 2007; pp. 1027–1035. [CrossRef]