processes

MDPI

Article

Operation Pattern Recognition of the Refrigeration, Heating and Hot Water Combined Air-Conditioning System in Building Based on Clustering Method

Yabin Guo [1], Jiangyan Liu [2], Changhai Liu [1], Jiayin Zhu [1], Jifu Lu [1] and Yuduo Li [1,*]

[1] School of Water Conservancy and Civil Engineering, Zhengzhou University, Zhengzhou 450001, China
[2] Key Laboratory of Low-Grade Energy Utilization Technologies and Systems, Chongqing University, Chongqing 400044, China
* Correspondence: yd_li@gs.zzu.edu.cn

Abstract: Air-conditioning system operation pattern recognition plays an important role in the fault diagnosis and energy saving of the building. Most machine learning methods need labeled data to train the model. However, the difficulty of obtaining labeled data is much greater than that of unlabeled data. Therefore, unsupervised clustering models are proposed to study the operation pattern recognition of the refrigeration, heating and hot water combined air-conditioning (RHHAC) system. Clustering methods selected in this study include K-means, Gaussian mixture model clustering (GMMC) and spectral clustering. Further, correlation analysis is used to eliminate the redundant characteristic variables of the clustering model. The operating data of the RHHAC system are used to evaluate the performance of proposed clustering models. The results show that clustering models, after removing redundant variables by correlation analysis, can also identify the defrosting operation mode. Moreover, for the GMMC model, the running time is reduced from 27.80 s to 10.04 s when the clustering number is 5. The clustering performance of the original feature set model is the best when the number of clusters of the spectral clustering model is two and three. The clustering hit rate is 98.99%, the clustering error rate is 0.58% and the accuracy is 99.42%.

Keywords: air conditioning system; pattern recognition; clustering; correlation analysis; defrosting operation mode

Citation: Guo, Y.; Liu, J.; Liu, C.; Zhu, J.; Lu, J.; Li, Y. Operation Pattern Recognition of the Refrigeration, Heating and Hot Water Combined Air-Conditioning System in Building Based on Clustering Method. *Processes* **2023**, *11*, 812. https://doi.org/10.3390/pr11030812

Academic Editor: Jean-Pierre Corriou

Received: 8 February 2023
Revised: 1 March 2023
Accepted: 3 March 2023
Published: 8 March 2023

1. Introduction

According to the report of the International Energy Agency, buildings currently account for 36% of global carbon dioxide emissions [1]. China has also put forward carbon peak and carbon neutral targets, so energy conservation and emission reduction in various energy industries will become increasingly important [2,3]. Building cooling, heating and hot water supply account for a large proportion of building energy [4,5]. Therefore, the combined air-conditioning system that can achieve cooling, heating and hot water production will help achieve further energy-saving and emission-reduction goals. Analyzing the operation data of a combined air-conditioning system and identifying the operation mode through data mining will further help to improve the energy-saving potential.

At present, many researchers have carried out research on the refrigeration, heating and hot water combined air-conditioning (RHHAC) system. Gong et al. [6] studied the combined chiller system, which can provide a sanitary hot water supply and air conditioning simultaneously. Zhao et al. [7] studied the heating performance of an air-source heat pump with water tank for thermal energy storage. Byrne et al. [8,9] carried out the experimental study of an air-source heat pump for simultaneous heating and cooling. The research mainly consists of two parts, including basic concepts and performance verification and dynamic behavior and the two-phase thermosiphon defrosting technique. In addition, the operating characteristics of the combined air-conditioning system under

different operating conditions have also been studied, and especially the start-up characteristics have been analyzed. In the hot-water mode, the exhaust pressure of the unit reached 3.7 Mpa [10]. However, there is less research on the further in-depth analysis of the operation data of the RHHAC system. In particular, data-mining methods have been used more and more widely in the field of air conditioning [11–13], so it is necessary to carry out data-mining research on the RHHAC system. The clustering method is an unsupervised machine learning method [14–16] that can analyze the data without labeling data and obtain valuable information [17]. Since labeled data are more difficult to obtain than unlabeled data, the application of clustering methods will have advantages [18]. At present, the widely used clustering methods mainly include K-means clustering [19–22], Gaussian mixture model clustering (GMMC) [23–25] and spectral clustering (SC) [26–28]. Xia et al. [29] proposed the cluster model based on the K-means method to find out the difference between the climate and the building thermal environment of three regions. David et al. [30] proposed a new climate classification method based on K-means, and the towns incorrectly classified were reduced between 0.128 and 7.702%. Nurseda et al. [31] combined the K-means method and the association rule mining method to study the balance of solar power generation. Islam et al. [32] proposed an enhanced brain tumor detection scheme based on the K-means method and principal component analysis. A framework using clustering methods for modeling time-varying operations in complex energy systems was developed. Different clustering methods in the domain of the objective function of two example operational optimization problems were compared [33]. Zhao et al. [34] developed the Gaussian mixture model to preprocess historical data to obtain steady-state measurements under various operating conditions. Further, the Gaussian mixture model was used to optimize the energy management of heterogeneous building neighborhoods [35]. Shen et al. [36] used the principal analysis and the Gaussian mixture model to establish a building type clustering model, which is computationally efficient and a more accurate reflection of the local urban microclimate. Wang et al. [37] proposed an optimal scheduling strategy for an electricity–hydrogen–gas–heat integrated energy system based on the spectral clustering method, and used the spectral clustering method to describe the uncertainty of the system. Guo et al. [38] proposed multi-view spectral clustering combined with simultaneous consensus graph learning and discretization. From the above research, it can be concluded that the data-mining method can carry out knowledge discovery from the operation data of the air-conditioning system. The data-mining method has important potential for air-conditioning system operation pattern recognition. However, there is no research on applying the clustering method to data analysis of the RHHAC system. Therefore, this study focuses on the application of the clustering method in the operation pattern recognition of the RHHAC system.

In addition, different feature variable sets also have an important impact on the complexity and running time of the model [39]. The correlation analysis approach can eliminate the redundant characteristic variables in the model [40–42]. Therefore, the correlation analysis method is used to eliminate the redundant characteristic variables in this study.

The main contribution of this paper is to establish cluster models of a RHHAC system using various clustering methods for system operation pattern recognition. The clustering performance of the model with different clustering methods, different numbers of clusters and different feature variable sets is analyzed. This method can be used for data mining and energy-saving pattern recognition of air-conditioning systems, and plays an important role in the fault identification and energy saving of air-conditioning systems.

The paper is organized as follows. Section 2 outlines the pattern recognition approach based on clustering. Section 3 describes the experiment and data introduction of the RHHAC system. Section 4 introduces the results and discussion of the clustering model in detail. Finally, the paper is concluded in Section 5.

2. Pattern Recognition Approach Based on Clustering

The pattern recognition approach proposed in this study is based on the unsupervised clustering algorithm, as shown in Figure 1. The unsupervised clustering method can identify valuable operating modes under the condition of unlabeled data. This section introduces the principles of three unsupervised clustering methods and the redundant variable elimination approach based on correlation analysis. Further, quantitative evaluation indexes are proposed to evaluate the performance of different clustering models.

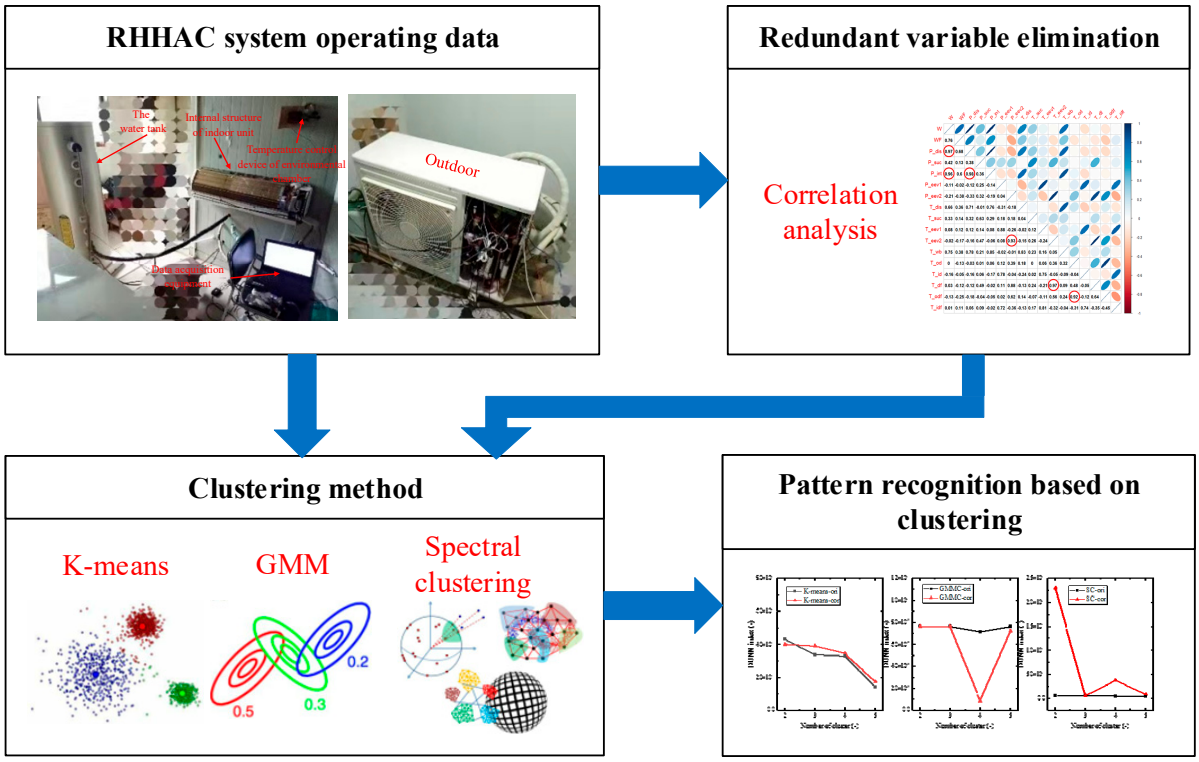

Figure 1. Pattern recognition approach flow chart based on clustering.

2.1. Redundant Variable Elimination Strategy Based on Correlation Analysis

The correlation coefficient reflects the degree of linear correlation among different variables. A positive correlation coefficient indicates a positive correlation among variables. A negative correlation coefficient indicates a negative correlation among variables. The larger the absolute value of the correlation coefficient, the stronger the correlation among the variables, and vice versa, the weaker the correlation. The Pearson simple correlation coefficient method is used in this study, which describes the correlation between two variables of different scales.

The correlation coefficient between the two variable data is defined as follows:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}} \quad (1)$$

In the above formula, r_{xy} is the value of the correlation coefficient, and the meaning of different values is shown in Figure 2:

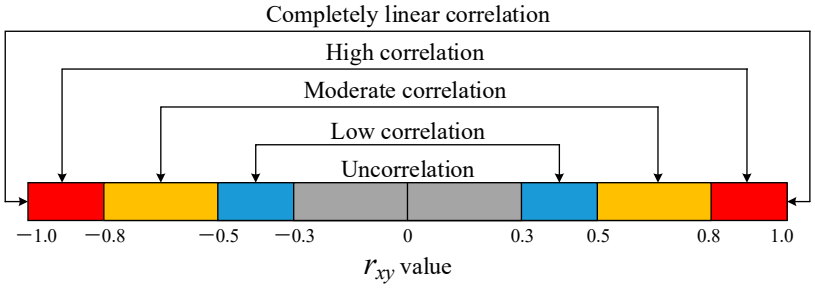

Figure 2. Correlation coefficient graph.

Variables with high correlation contain redundant information, and too many feature variables may increase the complexity of the model. Therefore, the correlation analysis method is selected to eliminate redundant variables, thereby reducing feature variables. When using the correlation analysis method to select feature variables, the choice of threshold is very important for the performance of eliminating redundant variables. If the threshold is too large, the performance of eliminating redundant variables will be poor. On the contrary, there will be a risk of loss of useful information. When the absolute value of r_{xy} is larger than 0.8, there is a high correlation between the two variables. However, in order to avoid excessive removal of feature variables and loss of effective information, when the absolute value of the correlation between the 2 feature variables is larger than 0.9, 1 of the 2 feature variables will be removed to achieve the purpose of reducing redundant variables.

2.2. Clustering Method

Cluster analysis is an important statistical analysis method for studying classification problems, and it is also an important algorithm in data mining. There are many implementation forms of cluster analysis, usually by calculating the sample distance in a multi-dimensional space. This research mainly uses the K-means method, Gaussian mixture model clustering method and spectral clustering method.

2.2.1. K-Means Approach

The K-means algorithm is a dynamic clustering algorithm, which belongs to the category of dynamic grouping. The basic idea is to randomly select K objects as the center of the initial K sets for a database containing N data objects. Then, the center distances of other samples in each set are calculated, and the set closest to the center sample is found. The average method is used to calculate the new cluster center after adjustment. If there is no change in the centers of two adjacent clusters, the sample clustering has been completed.

Calculation steps:

1. Select the number of clusters K.
2. Select K samples $C_1, C_2, \ldots, C_k$ as the initial cluster centers.
3. Calculate the distance d from other samples to the cluster center point, as shown in the formula:

$$d = \sqrt{\sum_{i=1}^{n} (x_i - C_{ji})^2} \quad C_j \in C \quad 1 < j < k \tag{2}$$

In the formula, x is a certain sample, and C_j is the center of a certain cluster.

4. According to the principle of being closest to the center point, all samples are classified into K categories.
5. Then, calculate the centroid of the cluster and use it as the new cluster center.
6. Repeat steps (3)–(5), and iterate until the cluster centers no longer change.

2.2.2. Gaussian Mixture Model Clustering Method

Gaussian mixture model clustering is a clustering method that uses multiple Gaussian models to represent data distribution. The clustering principle is as follows:

Assuming that the operating data sample of the combined air-conditioning system is $x_i(i = 1, 2, \ldots N)$, the Gaussian mixture model is shown in Equation (3):

$$P(x) = \sum_{k=1}^{K} \pi_k f(x_k \Big| \mu_k, \sum_k) \tag{3}$$

In the Gaussian mixture model, π, μ and $\sum$ need to be estimated, and Equation (3) can be transformed into Equation (4):

$$P(x \big| \pi, \mu, \sum) = \sum_{k=1}^{K} \pi_k f(x_k \Big| \mu_k, \sum_k) \tag{4}$$

The above three parameters can be estimated by the Expectation-Maximization (EM) method. The specific steps are as follows:

1. Set the initial values of π, μ and $\sum$.
2. Calculate the posterior probability $p(Z_{nk})$:

$$p(Z_{nk}) = \frac{\pi_k f(x_n | \mu_k, \sum_k)}{\sum_{j=1}^{K} \pi_j f(x_n \Big| \mu_j, \sum_j)} \tag{5}$$

3. Calculate the posterior probability $p(Z_{nk})$:

$$\mu_k = \frac{1}{f_k} \sum_{n=1}^{N} p(Z_{nk}) x_n \tag{6}$$

4. Find the maximum likelihood value of $\sum_k$:

$$\sum_k = \frac{1}{f_k} \sum_{n=1}^{N} p(Z_{nk})(x_n - \mu_k)(x_n - \mu_k)^T \tag{7}$$

5. Calculate the maximum likelihood function of π_k:

$$\pi_k = \frac{f_k}{f} \tag{8}$$

6. Perform iterative calculations on steps (2)–(5) until convergence.

2.2.3. Spectral Clustering Method

Spectral clustering (SC) is a clustering algorithm based on spectrogram theory. The main idea is to treat sample point data as points in space and connect the points with lines. The weight value is lower when the distance between two points is far, and vice versa. Then, the graph composed of all sample points is segmented. Furthermore, the weights in the sub-pictures are as high as possible after the segmentation process, and the weights among different sub-pictures are as low as possible, so as to achieve the clustering performance. The specific steps are shown in Figure 3.

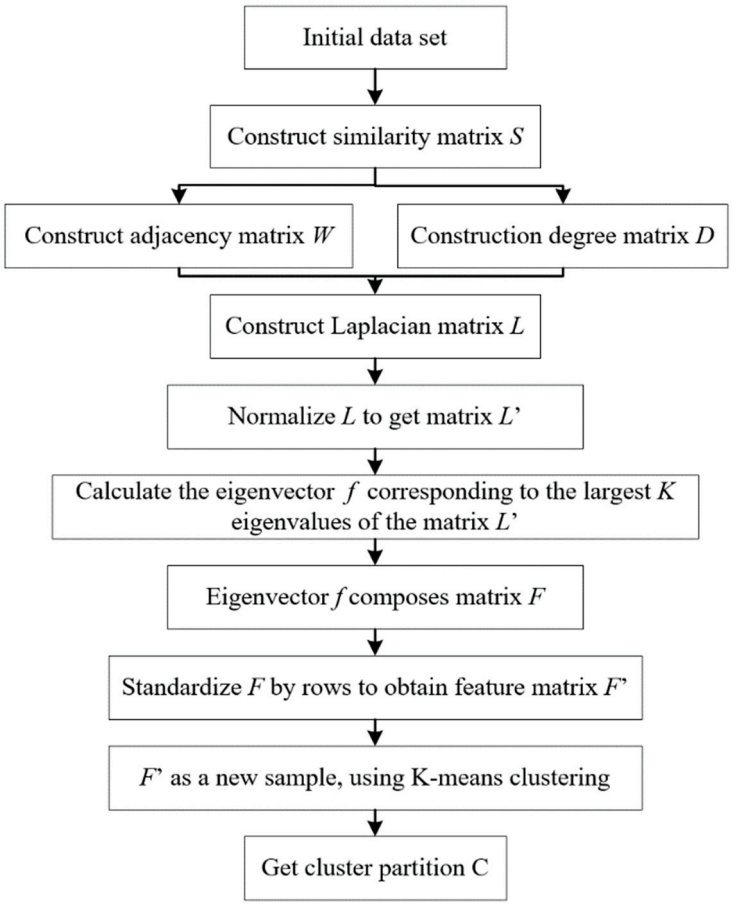

Figure 3. Schematic diagram of spectral clustering.

2.3. The Evaluation Index of Pattern Recognition Performance

In order to be able to evaluate the performance of clustering quantitatively, this study mainly uses two types of evaluation indicators, the DUNN index and the clustering accuracy index.

2.3.1. The DUNN Index

The DUNN index is widely used in the evaluation of clustering performance. This index is the shortest distance between any two samples in different clusters, divided by the maximum distance between two samples in any one cluster, and the calculation method is shown in Equation (9). The larger the DUNN index, the better the clustering performance.

$$\mathrm{DUNN} = \frac{\min\limits_{0<m\neq n<K}\left\{\min\limits_{\substack{\forall x_i\in\Omega_m\\ \forall x_j\in\Omega_n}}\left\{\left\|x_i-x_j\right\|\right\}\right\}}{\max\limits_{0<l\leq K}\max\limits_{\forall x_{i'},\forall x_{j'}\in\Omega_l}\left\{\left\|x_{i'}-x_{j'}\right\|\right\}} \tag{9}$$

2.3.2. Clustering Accuracy

In order to analyze the clustering performance more pertinently and accurately, this study proposes some new indicators to evaluate the clustering performance, mainly including the clustering hit rate, clustering error rate and accuracy. In the clustering result, the true category is Ψ_m, the expected cluster category is Ω_m, the number of correct samples in the expected cluster category is NT, and the number of incorrect samples in the expected cluster category is NF. The cluster hit rate C_h reflects the ratio of the number of clustered

correct samples in the expected cluster category, divided by the number of samples in the true category, and the calculation method is shown in Equation (10). In order to avoid a large number of wrong samples in the expected clustering category, the clustering accuracy index is also introduced. The accuracy C_p is the number of correct samples of the expected clustering category, divided by the number of samples of the expected clustering category. The calculation method is shown in Equation (11). In addition, the clustering error rate F represents the number of error samples in the expected cluster category, divided by the number of samples in the expected cluster category. The calculation method is shown in Equation (12). The higher the clustering hit rate and accuracy, the better. The lower the clustering error rate, the better.

$$C_h = \frac{NT}{N(\Psi_m)} \tag{10}$$

$$C_p = \frac{NT}{N(\Omega_m)} \tag{11}$$

$$F = \frac{NF}{N(\Omega_m)} \tag{12}$$

3. Experiment and Data Introduction

3.1. Introduction of Experimental Subjects

A RHHAC system is a system with refrigeration, heating, hot water and multiple complex operation modes. For example, a RHHAC system can produce hot water while cooling. Therefore, the structure of the RHHAC system has been improved compared with the ordinary air-conditioning system. The RHHAC system, mainly through two four-way reversing valves and two electronic expansion valves to realize the transformation of the refrigerant flow path, realizes the switching of various working conditions, such as cooling, cooling and hot water at the same time, in heating and hot-water modes. Figure 4 shows the schematic diagram of the RHHAC system structure and the refrigerant flow path layout under cooling mode conditions. The compressor type of the experimental system is QXA-C18B030, and the rated power is 1.5 kW. The type of the electronic expansion valve is DPF (TS1) 1.3C-01. The water tank capacity is 150 L. The RHHAC system implementation site diagram is shown in Figure 5. The RHHAC experimental system mainly includes an indoor unit, outdoor unit, heating water tank, environmental room and data acquisition device. The biggest feature of the RHHAC system is that it can meet the needs of users for hot water while realizing the cooling and heating functions.

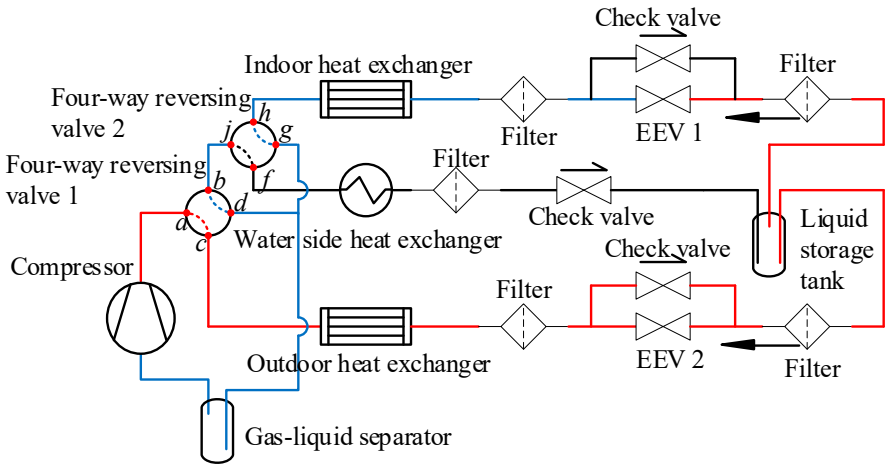

Figure 4. Schematic diagram of the RHHAC system.

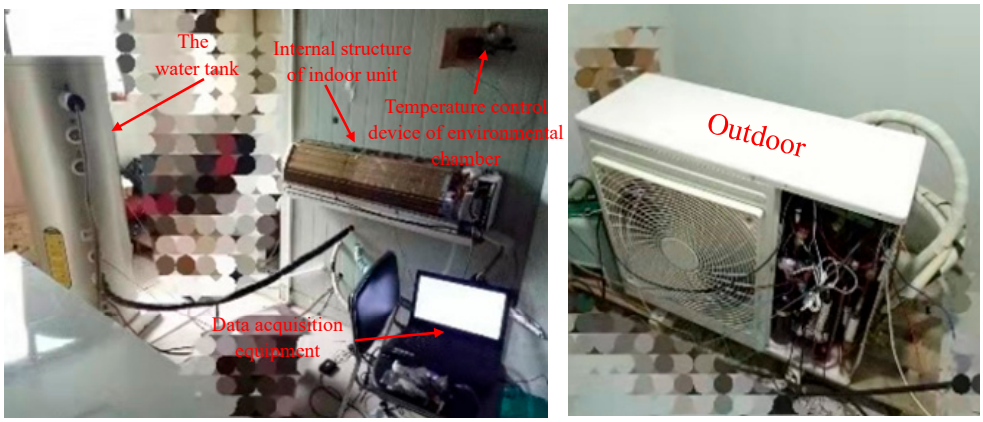

Figure 5. The RHHAC system implementation site diagram.

3.2. Data Collection and Processing

The experimental condition is that the outdoor temperature is 7 °C and the indoor temperature is 20 °C. Further, the initial temperature of the water tank is 30 °C., and the operating mode of the RHHAC system is the hot-water mode. The experiment collects a total of 50 variables in temperature, pressure and various control parameters. After removing invalid variables and some control parameters, 17 variables are retained, and the details of the selected variables are listed in Table 1. All sensors are installed by the unit itself. Table 2 lists the information of the temperature sensor and the pressure sensor. The data collection time interval is 1 s, and the collected data is transmitted to the computer through the host computer and stored. The experiment lasted 3.49 h, and 12,551 samples were collected. Each sample contains 17 selected variables.

Table 1. Details of collected experimental data.

No.	Variables	Unit	Maximum	Minimum	Average Value
1	Input power (W)	W	2072.60	352.40	1517.96
2	Power factor (WF)	-	0.97	0.47	0.94
3	Discharge pressure (P_{dis})	Mpa	3.69	0.47	2.31
4	Suction pressure (P_{suc})	Mpa	1.23	0.13	0.45
5	Intermediate pressure (P_{int})	Mpa	3.64	0.86	2.26
6	Pressure after electronic expansion valve 1 throttling (P_{eev1})	Mpa	3.18	0.19	0.67
7	Pressure after electronic expansion valve 2 throttling (P_{eev2})	Mpa	1.71	0.21	0.47
8	Discharge temperature (T_{dis})	°C	104.50	50.20	74.35
9	Suction temperature (T_{suc})	°C	20.80	−30.70	−0.48
10	Temperature after electronic expansion valve 1 (T_{eev1})	°C	48.30	−28.10	5.56
11	Temperature after electronic expansion valve 2 (T_{eev2})	°C	25.20	−22.50	−8.80
12	Water tank temperature (T_{wb})	°C	54.70	24.90	34.03
13	Outdoor temperature (T_{od})	°C	11.10	−0.10	3.67
14	Indoor temperature (T_{id})	°C	22.80	19.00	20.30
15	Defrost temperature (T_{df})	°C	17.10	−15.00	−5.23
16	Outdoor fan outlet temperature (T_{odf})	°C	21.43	0.02	3.74
17	Indoor fan outlet temperature (T_{idf})	°C	32.68	0.21	20.57

Table 2. Information and accuracy of sensors.

No.	Sensors	Type	Brand	Range	Accuracy
1	Temperature	Thermal resistance	UNIOHM	−30~120 °C	±1%
2	Pressure	Strain mode	Huadian	0~10 Mpa	±0.2%

4. Results and Discussion

4.1. Correlation Analysis Results

Figure 6 shows the results of the correlation analysis of the RHHAC system operating data. The lower left part of the figure is the specific correlation coefficient value, and the upper right part represents the correlation between variables in the form of a graph. The depth of the color represents the value of the correlation coefficient, and the deflection direction of the ellipse indicates that the variables are positively correlated or negatively correlated. By analyzing the correlation between the various variables, there are a total of 6 sets of variables with a correlation of more than 0.90. They mainly include W–P_{dis}, W–P_{int}, P_{int}–P_{dis}, T_{eev2}–P_{eev2}, T_{df}–T_{eev2} and T_{odf}–T_{od}. There is also overlap between these variable sets. Based on the consideration of eliminating more variables and better reflecting the operation of the system, the five variables of P_{dis}, P_{int}, P_{eev2}, T_{df} and T_{odf} are eliminated. The feature variable set after the correlation eliminates redundant variables (cor-feature) contains 12 variables, namely, W, WF, P_{suc}, P_{eev1}, T_{dis}, T_{suc}, T_{eev1}, T_{eev2}, T_{wb}, T_{od}, T_{id} and T_{idf}.

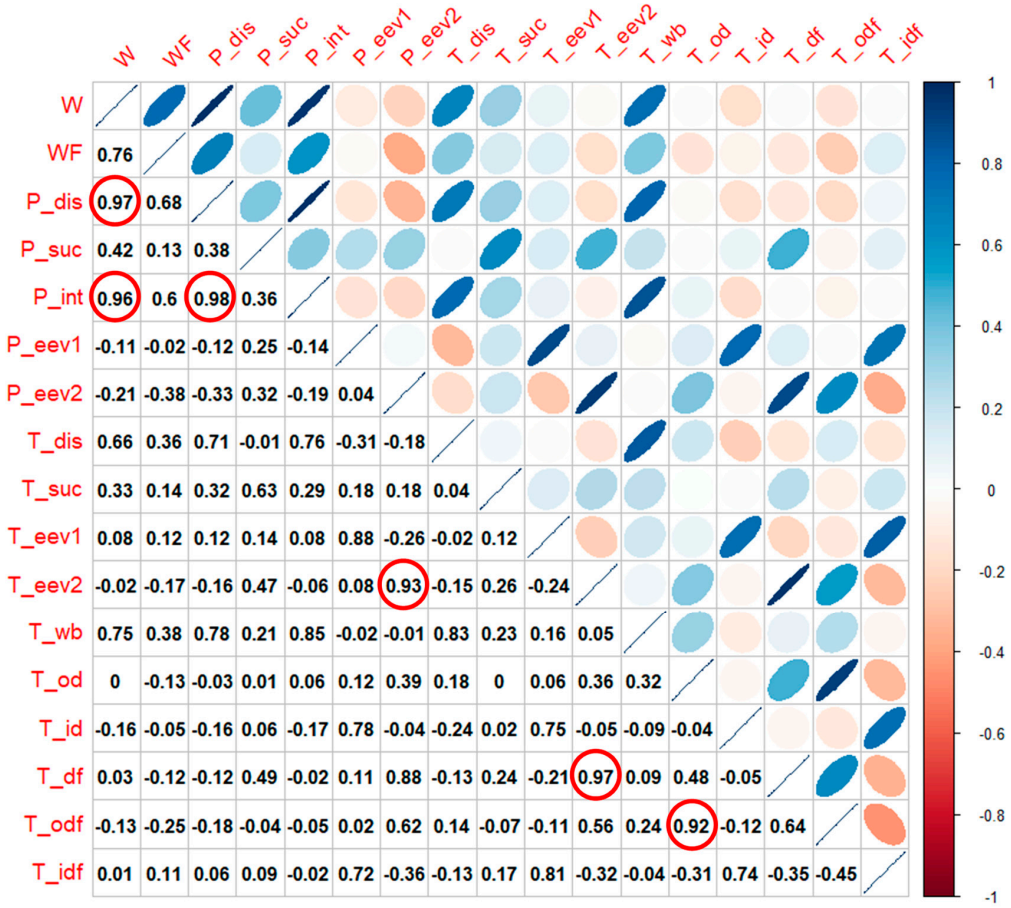

Figure 6. Result of correlation analysis of the RHHAC system operating data.

4.2. Clustering Result Analysis

Firstly, the original feature variable set (ori-feature) is used for cluster analysis. A total of 17 variables are included; detailed variable information is listed in Table 1.

The collected data are first used for K-means clustering analysis. The number of clusters is selected as two and three, and the results are shown in Figure 7. In order to present the clustering performance of high-dimensional data in a two-dimensional graph, the principal component result is used as the horizontal and vertical axis. The contribution of the abscissa is 30%, and the contribution of the ordinate is 25.5%. When clustering into two categories, it can be seen that the two categories are more distinct. Only in the middle three regions, the two categories have an intersection, and the data are divided

into two types, the upper left and the lower right. When the data are clustered into three categories, the boundaries of the three categories are also obvious, and there is not too much crossover between each category. There is not much overlap between the various categories. Through the comparative analysis of the two clustering results, it can be seen that cluster 1 is consistent. Cluster 2 and cluster 3, when the model clustered into three categories, correspond to cluster 2 when the model clustered into two categories.

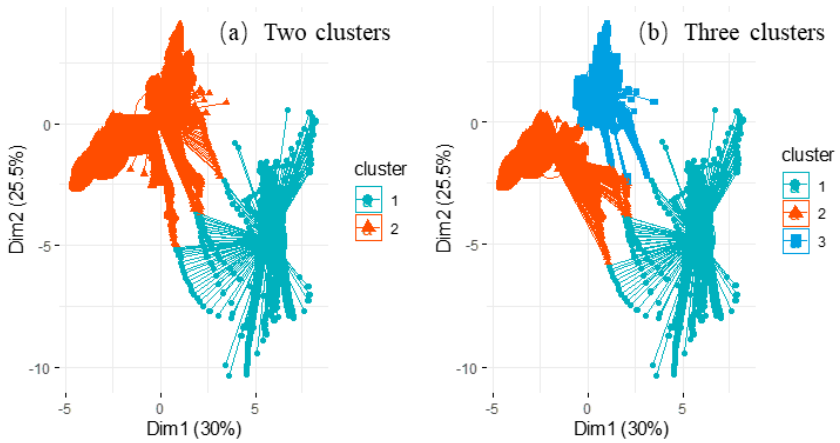

Figure 7. K-means clustering performance when numbers of clusters are two and three.

In order to further analyze the clustering results, each cluster is analyzed separately. The parameters that can better characterize the operation of the RHHAC system are selected, which mainly include the compressor discharge temperature, compressor suction temperature and defrost temperature. Figure 8 shows the results when the clustering is two categories. The figure shows the distribution of the operation data of two categories. Figure 9 shows the data distribution of cluster 1 when the number of clusters is 3. The other two categories have no clear knowledge information, so they are not displayed. First of all, it can be seen from Figure 8 that the data distributions of the two clusters have obvious differences. Cluster 1 compressor suction and discharge temperature fluctuate sharply, while cluster 2 compressor suction and discharge temperature change smoothly. The difference in defrosting temperature is even more obvious. The defrost temperature of cluster 1 changes regularly, which is worthy of further analysis. Comparing with the result of Figure 9, the changes in the two cases are similar. In addition, the changes in the suction and discharge temperature in cluster 1 of Figures 8 and 9 are also similar, which also illustrates that these two clusters are the same cluster and have the regular change period. Combining with the operating characteristics of the RHHAC system, it is found that cluster 1 is the defrosting process of the RHHAC system. It can also be seen from the figure that this cluster contains a total of four defrosting processes.

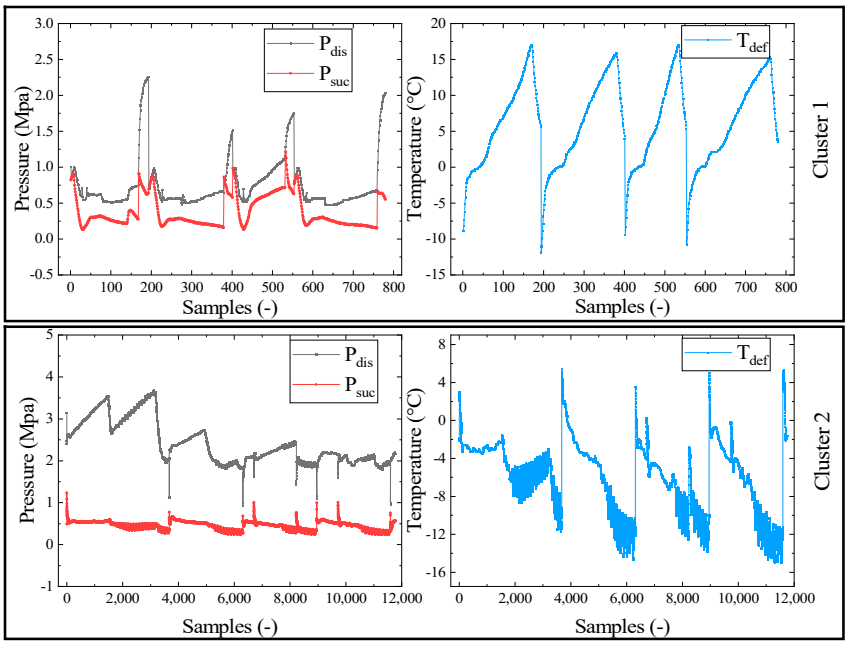

Figure 8. The K-means clustering result when the number of clusters is 2.

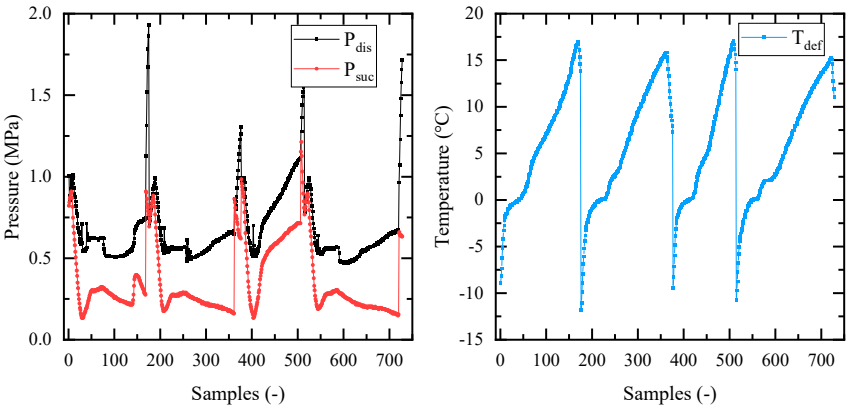

Figure 9. The K-means clustering result cluster 1 when the number of clusters is three.

4.3. Clustering Performance of Different Clustering Models

In order to carry out a more in-depth study on the clustering of the RHHAC system operating mode, this study uses three clustering methods to cluster the data of the RHHAC system, namely, K-means clustering, Gaussian mixture model clustering and spectral clustering. The cluster model data set includes the original feature variable set (ori) and the feature variable set after the correlation analysis removes redundant variables (cor). The number of clusters includes two, three, four and five, a total of four clustering situations.

4.3.1. K-Means Clustering Results

In order to analyze the clustering performance more intuitively, a clustering category diagram is used to show the clustering results of the RHHAC system data with different numbers of clusters. Figure 10 shows the comparison of the clustering results of the K-means algorithm. By analyzing the operating status of the system, the defrosting operation mode is artificially marked, as shown in the figure. The black line is the artificial marking state, the state 0.5 is the defrosting state, and the state −0.5 is the normal running state. The defrosting status marks in the results of other models are consistent with this. It can be seen from the figure that when the original feature variable set is used, there is a corresponding defrost category for different cluster numbers, which are cluster1, cluster1, cluster1 and

cluster5. The model can effectively cluster the defrost categories and has a good consistency. For the cor-feature set model, it is found that when the number of clusters is two and three, the clustering performance is poor, and the defrost category cannot be identified. When the number of clusters is four and five, the clustering performance is improved, and the defrosting operation modes can be identified, which are cluster 1 and cluster 4, respectively. These results show that the clustering method can identify the mode of defrosting operation. When the number of clusters is small, the clustering performance of the K-means model with the ori-feature set is better than the cor-feature set model.

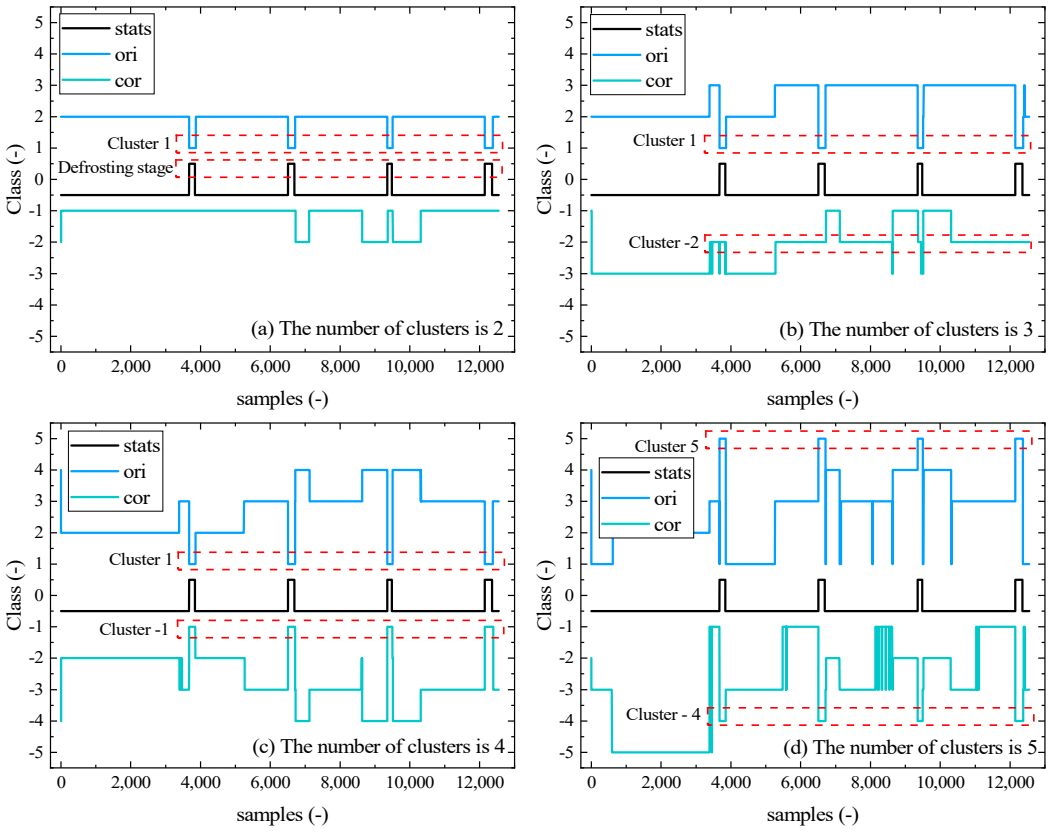

Figure 10. Comparison of the K-means clustering model results. (The "stats" represents the defrosting operation mode of the system. The "ori" represents clustering using original data sets. The "cor" represents the data set after removing redundant variables for clustering.).

4.3.2. Gaussian Mixture Model Clustering Results

Figure 11 shows the clustering results of the GMMC model. First, the model results of the ori-feature set are analyzed. When the number of clusters is 2, it can be found from the figure that cluster 2 clustered by GMMC is more consistent with the defrost category. Yet, it is obviously different from the result of the K-means algorithm, that is, the range of cluster 2 clustered by the GMMC algorithm exceeds the actual defrost category. From another perspective, the GMMC algorithm recognizes part of the normal operating state data as the defrosting category. Comparing the situation with other cluster numbers, there is a category corresponding to the defrost mode in the results of these models. The distribution of these categories is also relatively similar; these categories are cluster 2, cluster 3, cluster 3 and cluster 3. Then, the clustering performance of the cor-feature set is compared and analyzed. When the number of clusters is different, the clustering performance of the model is consistent with the results of the ori-feature set model. The corresponding clusters of defrosting mode obtained by clustering are cluster 2, cluster 3, cluster 3 and cluster 3. From these results, it can be concluded that for the GMMC model, removing redundant variables through correlation analysis will not affect the clustering performance of the model.

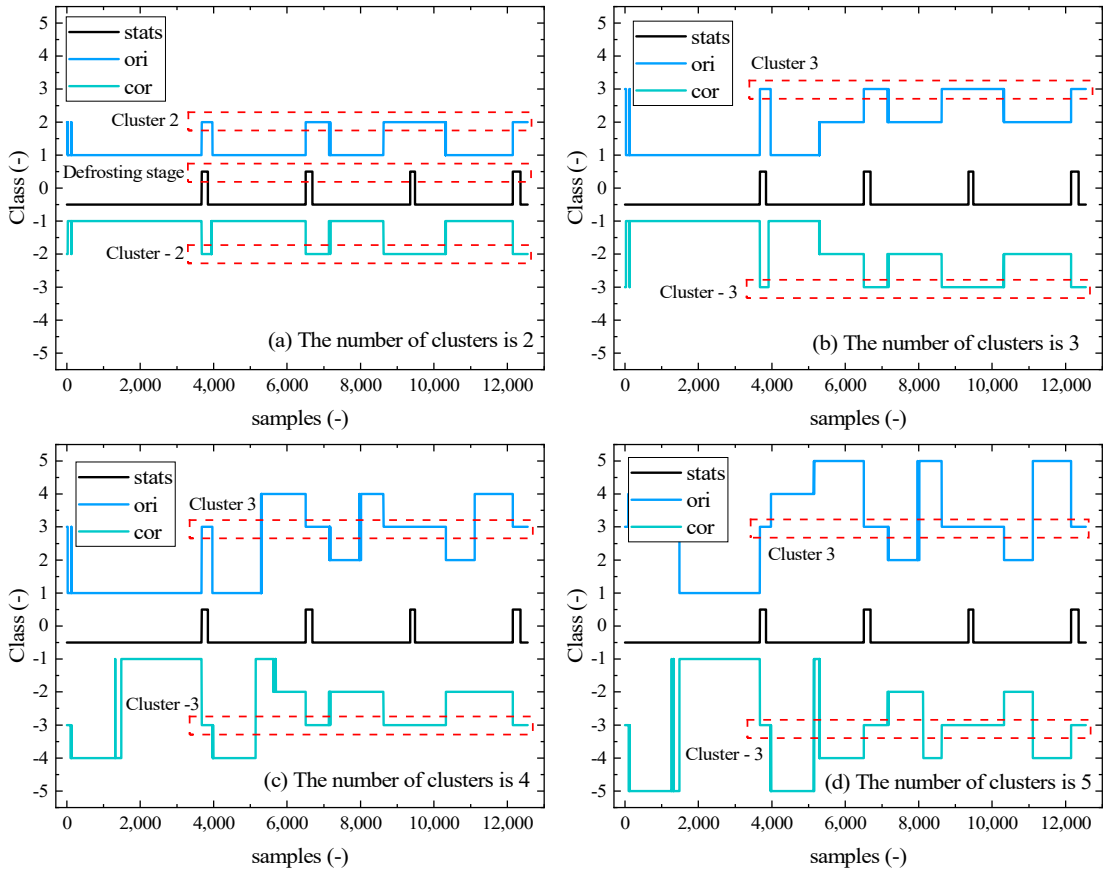

Figure 11. Comparison of the GMM clustering model results (The "stats" represents the defrosting operation mode of the system. The "ori" represents clustering using original data sets. The "cor" represents the data set after removing redundant variables for clustering.).

4.3.3. Spectral Clustering Results

Figure 12 shows the clustering results of the spectral clustering algorithm. For the clustering model established by the ori-feature set, when the number of clusters is two and three, cluster 2 and cluster 1 in the clustering result are consistent with the actual defrost category. It can be seen from the figure that no excessive operating data are identified as the defrost category. Yet, when the number of analysis clusters is four and five, category 2 and category 3 in the clustering result correspond to the actual defrost category. It can be seen from the figure that although defrosting can be identified, the clustering results contain more running data. Yet, when the number of clusters is four and five, cluster 2 and cluster 3 in the clustering result correspond to the actual defrost category. It can be seen from the figure that although the defrost category can be identified, the clustering result contains a lot of normal running data. For the cor-feature set model, when the number of clusters is three and four, the model can identify the defrosting operation mode accurately. Yet, when the other two cluster numbers are selected, the cluster performance is poor. Therefore, combining the clustering performance of the two feature sets, the best number of clusters is three.

4.4. Comparative Analysis of Clustering Performance

In order to analyze the clustering performance of each algorithm in different numbers of clusters more specifically, the DUNN index is used to evaluate the results of different numbers of clusters, as shown in Figure 13. For the K-means algorithm, as the number of clusters increases, the DUNN index gradually decreases, while for the GMMC and SC algorithms, the DUNN index fluctuates. In addition, for the K-means algorithm and the SC algorithm, the DUNN index of the cor-feature set model is basically higher than

that of the ori-feature set model. Yet, for the GMMC algorithm, the DUNN index of the cor-feature set model is lower than that of the ori-feature set model. The overall analysis found that the DUNN index of the model is higher when the number of clusters is small, which corresponds to the above clustering results. Yet, for the clustering models of different methods, the DUNN index cannot accurately evaluate the clustering performance of the model.

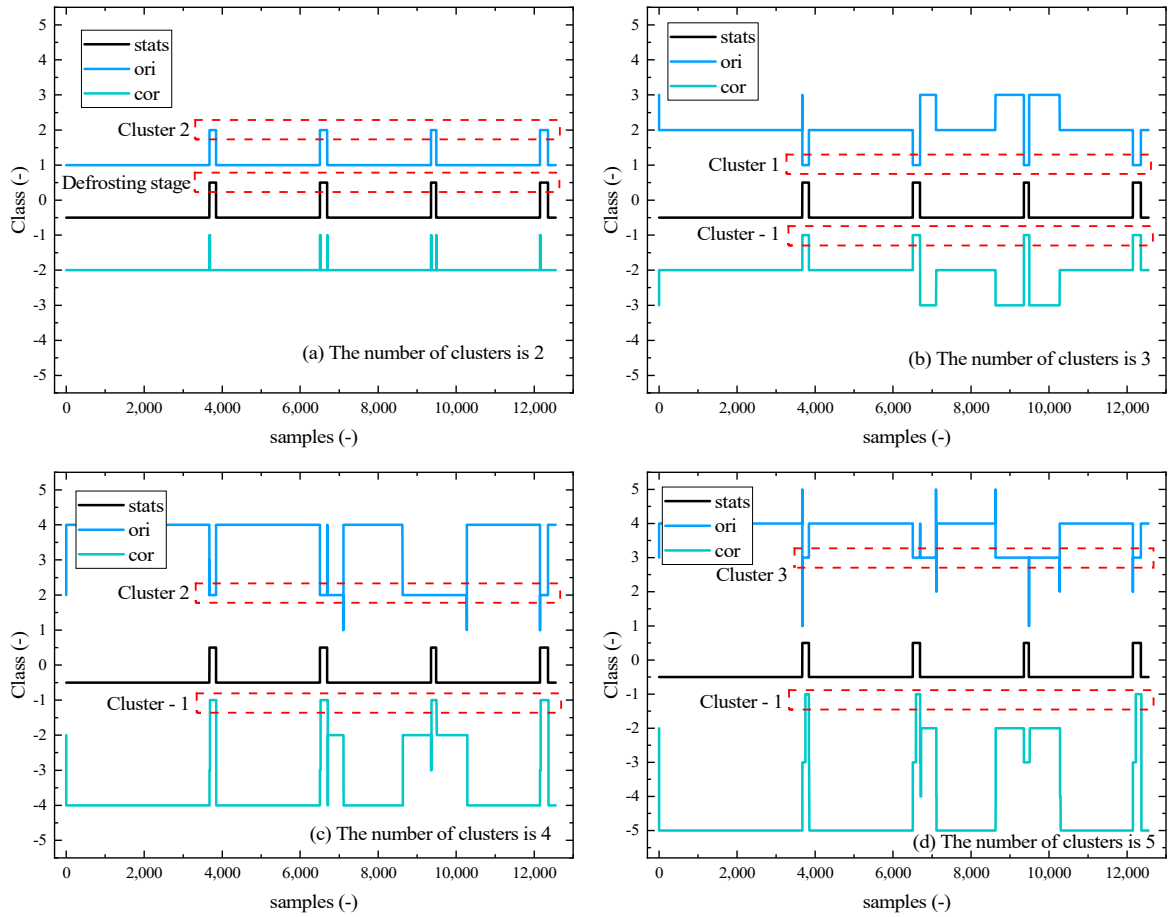

Figure 12. Comparison of the SC clustering model results (The "stats" represents the defrosting operation mode of the system. The "ori" represents clustering using original data sets. The "cor" represents the data set after removing redundant variables for clustering.).

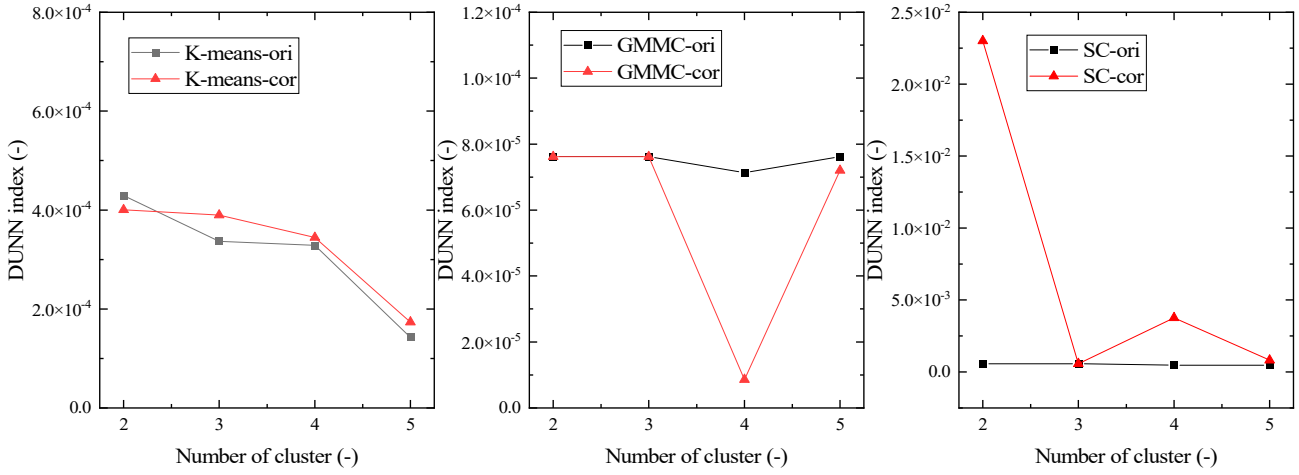

Figure 13. The DUNN index value under different numbers of clusters.

Table 3 lists the clustering accuracy rate of different clustering methods with different numbers of clusters. It can be seen from the overall results that the hit rate of clustering is relatively high. Among them, for the GMMC model with different numbers of clusters, the hit rate reached 100%, that is, the categories obtained by the cluster method completely included the defrosting operation mode data. Yet, from another aspect, the non-defrosting categories included in the clustering results are also worthy of attention. For the GMMC algorithm also, although the hit rate is higher, the error rate is also higher. When the number of clusters is 5, the clustering error rate reaches 77.94%, that is, the clustering result contains a large amount of running data. Further, this study also uses the clustering accuracy evaluation index, and it can be concluded that the model clustering accuracy is not high for the GMMC algorithm. Therefore, it is effective and accurate to use the three indicators of hit rate, error rate and accuracy to evaluate the clustering performance of the model.

Table 3. Performance of different clustering models.

Number of Clusters			2 Clusters	3 Clusters	4 Clusters	5 Clusters
K-means	ori	True	98.99%	98.99%	98.99%	98.99%
		False	11.79%	10.65%	11.11%	7.90%
		Precision	88.21%	89.35%	88.89%	92.10%
	cor	True	98.13%	93.53%	**98.99%**	98.99%
		False	93.58%	88.60%	**10.07%**	8.27%
		Precision	6.42%	11.40%	**89.93%**	91.73%
GMMC	ori	True	100%	100%	100%	100%
		False	77.33%	77.28%	77.38%	77.94%
		Precision	22.67%	22.72%	22.62%	22.06%
	cor	True	**100%**	**100%**	**100%**	**100%**
		False	**77.06%**	**76.83%**	**77.96%**	**77.92%**
		Precision	**22.94%**	**23.17%**	**22.04%**	**22.08%**
SC	ori	True	**98.99%**	**98.99%**	99.14%	98.99%
		False	**0.58%**	**0.58%**	73.70%	73.70%
		Precision	**99.42%**	**99.42%**	26.30%	26.30%
	cor	True	7.48%	98.99%	91.51%	46.47%
		False	10.34%	1.43%	7.02%	4.44%
		Precision	89.66%	98.57%	92.98%	95.56%

Combining the three evaluation indexes, we compare and analyze the performance of the ori-feature set and cor-feature set model under different methods. First, for the K-means method, the clustering performance of the cor-feature set model is better when the number of clusters is four. Then, for the GMMC method, the clustering performance of the cor-feature set model and the ori-feature set model is not significantly different. Finally, for the SC method, the clustering performance of the ori-feature set model is better. This result also shows that the clustering model still has a good clustering effect and greatly reduces the complexity of the model after eliminating redundant variables through correlation analysis.

Based on the above analysis, it is concluded that the clustering performance of the ori-feature set model is the best when the number of clusters of the SC algorithm model is two and three. The hit rate is 98.99%, the error rate is 0.58%, and the accuracy is 99.42%. When the number of clusters of the K-means model is five, the clustering performance of the ori-feature set is the second. For the GMMC model, its clustering performance is worse than the other two clustering methods.

4.5. The Clustering Model Running Time Analysis

Figure 14 shows the running time results of the model when the number of clusters is different. All models are run on a desktop computer with a CPU of Intel Core I7-6700 3.4 GHz, two memories of 8 G and a Windows 10 64-bit operating system. For the K-means algorithm, the running time of the model gradually increases as the number of clusters increases. For the GMMC method, when the ori-feature set is used, the running time of the model increases as the number of clusters increases. Yet, when the cor-feature set is used, the running time of the model is relatively stable. Compared with the ori-feature set model, the running time is greatly reduced. For the SC method, as the number of clusters increases, the running time of the model tends to decrease. There is little difference in the running time of models with different feature sets. When comparing and analyzing different clustering algorithms, the running time of the K-means algorithm is the shortest, which is less than 1 s. The running time of the SC algorithm is relatively long, more than 12 h, and the running time far exceeds the K-means algorithm and the GMMC algorithm. From the perspective of different feature sets, for the GMMC method, the running time of the cor-feature set model is greatly reduced. Especially when the number of clusters is 5, the running time of the model is reduced from 27.80 s to 10.04 s, a reduction of 63.88%. However, the running time is slightly increased under some other models.

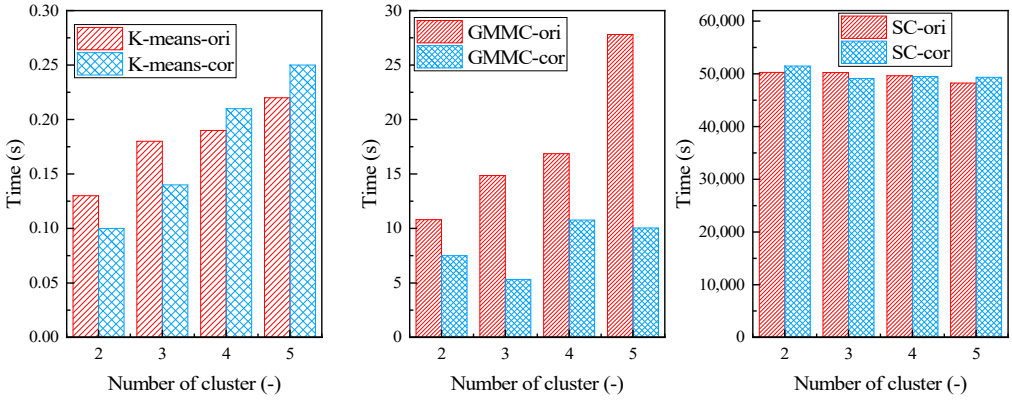

Figure 14. Comparison of the model running times with different numbers of clusters.

5. Conclusions

In this study, data mining is carried out based on the operating data of the RHHAC system. The following conclusions are drawn:

1. The operation data of the RHHAC system are analyzed by the clustering method. Different clustering methods can identify the defrosting operation mode.
2. Correlation analysis can eliminate redundant feature variables in the model. The cor-feature set model can also identify the defrosting operation mode under the condition of fewer feature variables. Some models are even better than the ori-feature set model. The running time of the model is also improved with different feature sets. Especially for the GMMC method when the number of clusters is 5, the model running time is reduced from 27.80 s to 10.04 s, which is a reduction of 63.88%.
3. The DUNN index cannot evaluate the clustering performance of the model very accurately. Analyzing the clustering performance evaluation indexes proposed in this study, the clustering performance of the ori-feature set model is the best when the number of clusters of the SC algorithm is two and three. The clustering hit rate is 98.99%, the clustering error rate is 0.58%, and the accuracy is 99.42%. When the number of clusters of the K-means model is five, the clustering performance of the ori-feature set is the second. For the GMMC model, its clustering performance is worse than the other two clustering methods.

The clustering model established in this study can effectively identify the operation mode of the RHHAC system. In the future, clustering will be further used to carry out relevant research on identifying energy-saving modes and fault modes of air-conditioning systems.

Author Contributions: Methodology, Y.G.; resources, C.L.; data curation, J.Z.; writing—original draft preparation, Y.G.; writing—review and editing, J.L. (Jiangyan Liu); visualization, C.L.; supervision, Y.L.; funding acquisition, J.L. (Jifu Lu) and Y.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the China Construction Seventh Bureau Technology Research Project (No. CSCEC7b-2015-Z-24) and the Science and Technology Department of Henan Province (No. 222102320051 and No. 222102320113).

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.

Nomenclature

C_h	cluster hit rate
C_j	center of a certain cluster
C_p	clustering accuracy
cor	feature variable set after the correlation analysis removes redundant variables
f	eigenvector
F	clustering error rate
K	number of clusters
ori	original feature variable set
P_{dis}	discharge pressure
P_{suc}	suction pressure
P_{int}	intermediate pressure
P_{eev1}	pressure after electronic expansion valve 1 throttling
P_{eev2}	pressure after electronic expansion valve 2 throttling
r_{xy}	value of the correlation coefficient
Ψ_m	true category
T_{dis}	discharge temperature
T_{suc}	suction temperature
T_{eev1}	temperature after electronic expansion valve 1
T_{eev2}	temperature after electronic expansion valve 2
T_{wb}	water tank temperature
T_{od}	outdoor temperature
T_{id}	indoor temperature
T_{df}	defrost temperature
T_{odf}	outdoor fan outlet temperature
T_{idf}	indoor fan outlet temperature
	Greek symbols
π	weighting coefficient
μ	the mean
$\sum$	covariance matrix
Ω_m	expected cluster category

References

1. Dai, B.; Tong, Y.; Hu, Q.; Chen, Z. Characteristics of thermal stratification and its effects on HVAC energy consumption for an atrium building in south China. *Energy* **2022**, *249*, 123425. [CrossRef]
2. Jin, B. Research on performance evaluation of green supply chain of automobile enterprises under the background of carbon peak and carbon neutralization. *Energy Rep.* **2021**, *7*, 594–604. [CrossRef]

3. Fang, K.; Li, C.; Tang, Y.; He, J.; Song, J. China's pathways to peak carbon emissions: New insights from various industrial sectors. *Appl. Energy* **2022**, *306*, 118039. [CrossRef]
4. Rogers, A.; Guo, F.; Rasmussen, B. A review of fault detection and diagnosis methods for residential air conditioning systems. *Build. Environ.* **2019**, *161*, 106236. [CrossRef]
5. Fan, C.; Liu, Y.; Liu, X.; Sun, Y.; Wang, J. A study on semi-supervised learning in enhancing performance of AHU unseen fault detection with limited labeled data. *Sustain. Cities Soc.* **2021**, *70*, 102874. [CrossRef]
6. Gong, G.; Chen, F.; Su, H.; Zhou, J. Thermodynamic simulation of condensation heat recovery characteristics of a single stage centrifugal chiller in a hotel. *Appl. Energy* **2012**, *91*, 326–333. [CrossRef]
7. Zhao, X.; Long, E.; Zhang, Y.; Liu, Q.; Jin, Z.; Liang, F. Experimental Study on Heating Performance of Air—Source Heat Pump with Water Tank for Thermal Energy Storage. *Procedia Eng.* **2017**, *205*, 2055–2062. [CrossRef]
8. Byrne, P.; Miriel, J.; Lenat, Y. Experimental study of an air-source heat pump for simultaneous heating and cooling—Part 2: Dynamic behaviour and two-phase thermosiphon defrosting technique. *Appl. Energy* **2011**, *88*, 3072–3078. [CrossRef]
9. Byrne, P.; Miriel, J.; Lenat, Y. Experimental study of an air-source heat pump for simultaneous heating and cooling—Part 1: Basic concepts and performance verification. *Appl. Energy* **2011**, *88*, 1841–1847. [CrossRef]
10. Pei, X.; Guo, Y.; Guo, Y.; Li, R.; Yang, J.; Mao, M. Research on the characteristics of the refrigeration, heating and hot water combined air-conditioning system. *Int. J. Refrig.* **2021**, *130*, 150–160. [CrossRef]
11. Zhou, Z.; Chen, H.; Li, G.; Zhong, H.; Zhang, M.; Wu, J. Data-driven fault diagnosis for residential variable refrigerant flow system on imbalanced data environments. *Int. J. Refrig.* **2021**, *125*, 34–43. [CrossRef]
12. Guo, Y.; Chen, H. Fault diagnosis of VRF air-conditioning system based on improved Gaussian mixture model with PCA approach. *Int. J. Refrig.* **2020**, *118*, 1–11. [CrossRef]
13. Bai, X.; Zhang, M.; Jin, Z.; You, Y.; Liang, C. Fault detection and diagnosis for chiller based on feature-recognition model and Kernel Discriminant Analysis. *Sustain. Cities Soc.* **2022**, *79*, 103708. [CrossRef]
14. Li, G.; Hu, Y. Improved sensor fault detection, diagnosis and estimation for screw chillers using density-based clustering and principal component analysis. *Energy Build.* **2018**, *173*, 502–515. [CrossRef]
15. An, J.; Yan, D.; Hong, T. Clustering and statistical analyses of air-conditioning intensity and use patterns in residential buildings. *Energy Build.* **2018**, *174*, 214–227. [CrossRef]
16. Nweye, K.; Nagy, Z. MARTINI: Smart meter driven estimation of HVAC schedules and energy savings based on Wi-Fi sensing and clustering. *Appl. Energy* **2022**, *316*, 118980. [CrossRef]
17. Afaifia, M.; Djiar, K.A.; Bich-Ngoc, N.; Teller, J. An energy consumption model for the Algerian residential building's stock, based on a triangular approach: Geographic Information System (GIS), regression analysis and hierarchical cluster analysis. *Sustain. Cities Soc.* **2021**, *74*, 103191. [CrossRef]
18. Gilanifar, M.; Wang, H.; Cordova, J.; Ozguven, E.E.; Strasser, T.I.; Arghandeh, R. Fault classification in power distribution systems based on limited labeled data using multi-task latent structure learning. *Sustain. Cities Soc.* **2021**, *73*, 103094. [CrossRef]
19. Molokomme, D.; Chabalala, C.; Bokoro, P. Enhancement of Advanced Metering Infrastructure Performance Using Unsupervised K-Means Clustering Algorithm. *Energies* **2021**, *14*, 2732. [CrossRef]
20. Yu, W.; Zhao, F.; Yang, W.; Xu, H. Integrated analysis of CFD simulation data with K-means clustering algorithm for soot formation under varied combustion conditions. *Appl. Therm. Eng.* **2019**, *153*, 299–305. [CrossRef]
21. Wang, K.; Qi, X.; Liu, H.; Song, J. Deep belief network based k-means cluster approach for short-term wind power forecasting. *Energy* **2018**, *165*, 840–852. [CrossRef]
22. Jiménez Torres, M.; Bienvenido-Huertas, D.; May Tzuc, O.; Bassam, A.; Ricalde Castellanos, L.J.; Flota-Bañuelos, M. Assessment of climate change's impact on energy demand in Mexican buildings: Projection in sin-gle-family houses based on Representative Concentration Pathways. *Energy Sustain. Dev.* **2023**, *72*, 185–201. [CrossRef]
23. Jin, H.; Shi, L.; Chen, X.; Qian, B.; Yang, B.; Jin, H. Probabilistic wind power forecasting using selective ensemble of finite mixture Gaussian process regression models. *Renew. Energy* **2021**, *174*, 1–18. [CrossRef]
24. Qureshi, M.; Ghiaus, C.; Ahmad, N. A blind event-based learning algorithm for non-intrusive load disaggregation. *Int. J. Electr. Power Energy Syst.* **2021**, *129*, 106834. [CrossRef]
25. Yi, D.H.; Kim, D.W.; Park, C.S. Prior selection method using likelihood confidence region and Dirichlet process Gaussian mixture model for Bayesian inference of building energy models. *Energy Build.* **2020**, *224*, 110293. [CrossRef]
26. Gao, S.; Hu, B.; Xie, K.; Niu, T.; Li, C.; Yan, J. Spectral clustering based demand-oriented representative days selection method for power system expansion planning. *Int. J. Electr. Power Energy Syst.* **2021**, *125*, 106560. [CrossRef]
27. Pacella, M.; Papadia, G. Fault Diagnosis by Multisensor Data: A Data-Driven Approach Based on Spectral Clustering and Pairwise Constraints. *Sensors* **2020**, *20*, 7065. [CrossRef]
28. Li, P.-H.; Pye, S.; Keppo, I. Using clustering algorithms to characterise uncertain long-term decarbonisation pathways. *Appl. Energy* **2020**, *268*, 114947. [CrossRef]
29. Xia, B.; Han, J.; Zhao, J.; Liang, K. Technological adaptation zone of passive evaporative cooling of China, based on a clustering analysis. *Sustain. Cities Soc.* **2021**, *66*, 102564. [CrossRef]
30. Bienvenido-Huertas, D.; Marín-García, D.; Carretero-Ayuso, M.J.; Rodríguez-Jiménez, C.E. Climate classification for new and restored buildings in Andalusia: Analysing the current regulation and a new approach based on k-means. *J. Build. Eng.* **2021**, *43*, 102829. [CrossRef]

31. Yürüşen, N.Y.; Uzunoğlu, B.; Talayero, A.P.; Estopiñán, A.L. Apriori and K-Means algorithms of machine learning for spatio-temporal solar generation balancing. *Renew. Energy* **2021**, *175*, 702–717. [CrossRef]
32. Islam, M.K.; Ali, M.S.; Miah, M.S.; Rahman, M.M.; Alam, M.S.; Hossain, M.A. Brain tumor detection in MR image using superpixels, principal component analysis and template based K-means clustering algorithm. *Mach. Learn. Appl.* **2021**, *5*, 100044. [CrossRef]
33. Teichgraeber, H.; Brandt, A.R. Clustering methods to find representative periods for the optimization of energy systems: An initial framework and comparison. *Appl. Energy* **2019**, *239*, 1283–1293. [CrossRef]
34. Zhao, T.; Li, J.; Wang, P.; Yoon, S.; Wang, J. Improvement of virtual in-situ calibration in air handling unit using data preprocessing based on Gaussian mixture model. *Energy Build.* **2022**, *256*, 111735. [CrossRef]
35. Shafiullah, D.; Vergara, P.P.; Haque, A.; Nguyen, P.; Pemen, A. Gaussian Mixture Based Uncertainty Modeling to Optimize Energy Management of Heterogeneous Building Neighborhoods: A Case Study of a Dutch University Medical Campus. *Energy Build.* **2020**, *224*, 110150. [CrossRef]
36. Shen, P.; Liu, J.; Wang, M. Fast generation of microclimate weather data for building simulation under heat island using map capturing and clustering technique. *Sustain. Cities Soc.* **2021**, *71*, 102954. [CrossRef]
37. Wang, Z.; Hu, J.; Liu, B. Stochastic optimal dispatching strategy of electricity-hydrogen-gas-heat integrated energy system based on improved spectral clustering method. *Int. J. Electr. Power Energy Syst.* **2021**, *126*, 106495. [CrossRef]
38. Zhong, G.; Shu, T.; Huang, G.; Yan, X. Multi-view spectral clustering by simultaneous consensus graph learning and discretization. *Knowl.-Based Syst.* **2022**, *235*, 107632. [CrossRef]
39. Hu, M.; Ge, D.; Telford, R.; Stephen, B.; Wallom, D.C. Classification and characterization of intra-day load curves of PV and non-PV households using interpretable feature extraction and feature-based clustering. *Sustain. Cities Soc.* **2021**, *75*, 103380. [CrossRef]
40. Wang, J.; Li, G.; Chen, H.; Liu, J.; Guo, Y.; Hu, Y.; Li, J. Liquid floodback detection for scroll compressor in a VRF system under heating mode. *Appl. Therm. Eng.* **2017**, *114*, 921–930. [CrossRef]
41. Guo, Y.; Li, G.; Chen, H.; Wang, J.; Guo, M.; Sun, S.; Hu, W. Optimized neural network-based fault diagnosis strategy for VRF system in heating mode using data mining. *Appl. Therm. Eng.* **2017**, *125*, 1402–1413. [CrossRef]
42. Zhang, Q.; Xu, D.; Zhou, D.; Yang, Y.; Rogora, A. Associations between urban thermal environment and physical indicators based on meteorological data in Foshan City. *Sustain. Cities Soc.* **2020**, *60*, 102288. [CrossRef]