


## Article

# Clustering Method Comparison for Rural Occupant's Behavior Based on Building Time-Series Energy Data

Xiaodong Liu, Shuming Zhang, Xiaohan Wang, Rui Wu, Junqi Yang, Hong Zhang \*, Jianing Wu  and Zhixin Li

School of Architecture, Tsinghua University, Beijing 100084, China; liuyan6064331@tsinghua.edu.cn (X.L.); zhangsm20@mails.tsinghua.edu.cn (S.Z.); xh-wang22@mails.tsinghua.edu.cn (X.W.); wu-r22@mails.tsinghua.edu.cn (R.W.); yang-jq23@mails.tsinghua.edu.cn (J.Y.); wujn21@mails.tsinghua.edu.cn (J.W.); lizhixin22@mails.tsinghua.edu.cn (Z.L.)

\* Correspondence: zhanghong@tsinghua.edu.cn

**Abstract:** The purpose of this research is to compare clustering methods and pick up the optimal clustered approach for rural building energy consumption data. Research undertaken so far has mainly focused on solving specific issues when employing the clustered method. This paper concerns Yushan island resident's time-series electricity usage data as a database for analysis. Fourteen algorithms in five categories were used for cluster analysis of the basic data sets. The result shows that Km\_Euclidean and Km\_shape present better clustering effects and fitting performance on continuous data than other algorithms, with a high accuracy rate of 67.05% and 65.09%. Km\_DTW is applicable to intermittent curves instead of continuous data with a low precision rate of 35.29% for line curves. The final conclusion indicates that the K-means algorithm with Euclidean distance calculation and the k-shape algorithm are the two best clustering algorithms for building time-series energy curves. The deep learning algorithm can not cluster time-series-building electricity usage data under default parameters in high precision.

**Keywords:** cluster analysis; rural building; time-series electricity; carbon emission; energy efficiency



**Citation:** Liu, X.; Zhang, S.; Wang, X.; Wu, R.; Yang, J.; Zhang, H.; Wu, J.; Li, Z. Clustering Method Comparison for Rural Occupant's Behavior Based on Building Time-Series Energy Data. *Buildings* **2024**, *14*, 2491. <https://doi.org/10.3390/buildings14082491>

Academic Editor: Francesco Nocera

Received: 17 July 2024

Revised: 9 August 2024

Accepted: 11 August 2024

Published: 12 August 2024



**Copyright:** © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

### 1.1. Background

With the ongoing progression of the economy, global energy consumption levels are steadily rising, with building energy usage constituting approximately 40% of total energy consumption—a figure that continues to grow [1]—contributing to over 30% of CO<sub>2</sub> emissions [2].

To mitigate the current reliance on fossil fuels and, consequently, lessen the adverse effects on the global climate, attention has shifted towards clean energy sources such as solar, wind, and geothermal energy. Research indicates that the installed capacity of renewable energy was around 2800 GW in 2020 [3], increasing to 3064 GW in 2021 [4]. Furthermore, it is suggested that achieving an installed capacity of 27,700 GW of renewable energy by 2050 could help limit global temperature rise to no more than 1.5 degrees Celsius by that year [5].

Reducing the energy consumption tied to building operations is vital for reaching sustainability objectives. Effective energy management is essential for enhancing energy efficiency and minimizing both total energy usage and operational costs. Concurrently, sophisticated automatic control technologies have been developed to manage energy consumption with precision, contributing to the resolution of global energy issues. Nevertheless, despite the proliferation of energy-saving technologies and policies, applying advanced technologies and enacting policies still necessitates human intervention. Research has shown that human behavior significantly influences building energy consumption [6].

Data mining (DM) has emerged as a powerful method for uncovering patterns in building operation energy consumption data, garnering significant attention in recent years.

Unlike traditional statistical or physical principle-based approaches, DM excels at processing massive datasets, uncovering potentially valuable and previously unknown information, and requiring less domain expertise [7]. One of the key insights gained through DM techniques is load profiling, which involves grouping temporal subsequences of measured electricity data to discern typical electricity consumption patterns in buildings [8]. However, these raw electricity consumption patterns can be challenging to interpret [7]. Thus, the interpretation of clustering results, referred to as “knowledge discovery”, becomes an attractive and valuable step, making DM techniques more practical for real-world applications.

### 1.2. Literature Review

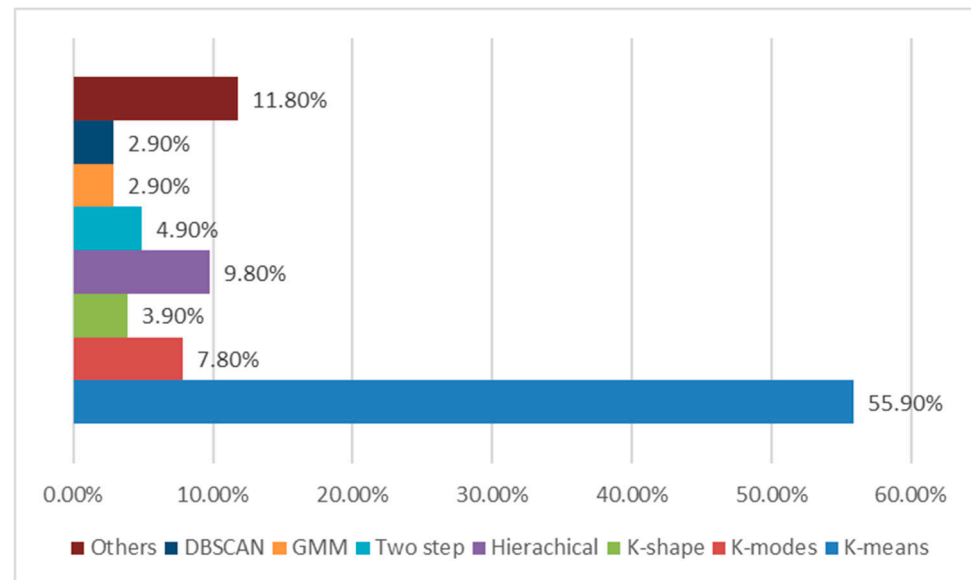
Unsupervised learning techniques are highly effective in discerning patterns in building electricity consumption from operational data [9]. Recognizing typical electricity load patterns (TELPs) is crucial for comprehending the characteristics of daily electricity load profiles (DELPs) in buildings [10]. Clustering, a commonly employed unsupervised learning method, is extensively used to extract building electricity load patterns by identifying inherent patterns within datasets [11]. Clustering algorithms utilized for analyzing electricity load profiles can be categorized into partition methods, hierarchical methods, density-based methods, and model-based methods [12].

For instance, Ma et al. utilized the partitioning around medoids algorithm to identify typical daily heating load profiles in higher education buildings, using the Pearson correlation coefficient instead of Euclidean distance to measure dissimilarity between cluster heating load profiles [13]. Agglomerative hierarchical clustering with combined dissimilarity measures was employed to discover electricity load profiles in two university library buildings [14]. These profiles are more likely to be grouped into the same cluster, with the K-means algorithm being a prominent example of this method [15]. Additionally, K-medoids and hierarchical clustering are significant components of this approach. Some recently proposed clustering methods also depend on Euclidean distance. Ref. [16] used cluster analysis of simulated energy consumption data from 134 US LEED NC office buildings to classify them into high, medium, and low energy use intensity clusters, showing that lower energy use is primarily due to reduced process and heating loads. Reference [17] identifies key variables influencing energy consumption in Singaporean office buildings using k-means clustering, highlighting gross floor area, non-air-conditioning energy consumption, chiller efficiency, and chiller capacity. A kRNN-LSTM deep learning framework integrated k-means algorithm was used for predicting and optimizing building energy management, achieving 94% accuracy using smart meter data [18]. Similarly, ref. [19] introduced a cluster-based aggregate forecasting method using k-medoids and the additive Gaussian process to improve residential load prediction accuracy. Ref. [20] investigated a framework using hierarchical clustering to identify inefficient rural US homes for energy efficiency improvements.

However, these traditional algorithms are inadequate for capturing temporal variations between data objects, which is a significant limitation for clustering occupant behavior patterns that are strongly time-dependent. Consequently, recent years have seen an increased interest among researchers in utilizing time-series clustering algorithms, such as the K-shape clustering algorithm, to enhance clustering accuracy [21]. As illustrated in Figure 1, over 55% of researchers employ the K-means clustering method. In contrast, approximately 8% use the K-mode clustering method, around 4% use the K-shape clustering method, and about 10% use hierarchical clustering methods. The remaining 22% of researchers utilize various other clustering algorithms [22].

The k-means algorithm, a classic partitioning clustering method, is frequently utilized in data mining literature due to its ease of implementation and high efficiency [11]. Ref. [23] compared the k-means, bisecting k-means, and Gaussian mixture model algorithms and found that k-means was the most suitable for analyzing building electricity load patterns in a dataset comprising 1910 residential and 1919 non-residential buildings. Ref. [24] pro-

posed an improved k-means clustering method that integrates optimal initial cluster centers with principal component analysis to enhance convergence speed using large-scale smart meter data. Additionally, ref. [25] used the k-means algorithm to identify daily heating electricity load profiles of 139 Danish dwellings, revealing two main clusters: one for weekday profiles and another for weekend profiles.



**Figure 1.** Percentage condition for various clustered usage in the literature.

However, for large time-series datasets with high dimensionality (24 or higher), some clustering algorithms, including k-means, become impractical and may not be suitable for grouping similar electricity load profiles. This issue is known as the “curse of dimensionality” [26]. Dynamic time warping was developed to measure the similarity between time series to address this issue, but it can be computationally intensive.

To mitigate these issues, researchers have developed strategies to reduce data dimensionality. One approach is feature definition, which involves describing each electricity load profile with a limited number of expert-defined features, avoiding additional parameters. Ref. [27] defined three load shape parameters extracted from the raw time-series data: the peak-base load ratio, working/nonworking day load ratio, and on-hour duration. Ref. [28] defined seven statistical features—mean, standard deviation, skewness, kurtosis, chaos, energy, and periodicity—to represent the raw time-series, and then applied k-means clustering. Ref. [29] divided a day into four periods: overnight, breakfast, daytime, and evening, and calculated the relative average electricity consumption for each period. Ref. [24] improved the K-means algorithm into shape-based clustering to identify electricity consumption patterns from residential smart meter data. Ref. [30] investigated a clustering method of k-means for smart meter electricity demand data, finding significant variability in household consumption patterns that challenge standard assumptions and highlight important implications for energy policy and demand response programs. Ref. [31] developed a novel symbolic hierarchical clustering method to cluster the building operation patterns. In other words, some advanced deep-learning approaches were also developed for the purpose of data mining [32]. Ref. [33] developed a Monte Carlo-based model coupled with the k-means algorithm, using real-time occupancy data to improve building energy simulation performance, showing significant load prediction improvements over fixed schedules. Compared to raw time series, these studies have demonstrated that feature-based clustering can enhance clustering performance while reducing time and computational costs.

In addition to clustering methods, association rule mining algorithms have been successfully applied to identify energy-inefficient appliance usage behaviors [34], window usage behaviors [35], energy-inefficient lighting system usage behaviors [36], and the impacts of occupants on residential electricity consumption [37]. These algorithms have also been used to detect sensor faults, device faults, and control strategies in building energy systems. For instance, Yu et al. identified an energy-inefficient exhaust fan control strategy and two air handling unit faults in an HVAC system using association rule mining algorithms based on operational data [38]. Recent research has shown that association rule mining methods can also uncover dynamic operation patterns in building energy systems. For example, a temporal association rule mining method [39] and a progressive pattern mining method [40] have been proposed to detect energetic anomalies in HVAC systems. Ref. [41] combined the analytic hierarchy process manner into ARM route archiving building energy systems post evaluation. Ref. [42] integrated anomaly detection and dynamic energy performance evaluation functions into one ARM workflow. Ref. [43] adopted the ARM method and successfully investigated the dynamic relationship between building patterns and the people moving mode, both pre-pandemic and during the pandemic. The results showed that the size of the floor and the number of rooms have a positive impact on higher occupancy levels. Ref. [44] improved the Apriori method, finding occupants' activities could lead to enhanced pollutant concentrations within 2 h in residential buildings.

According to the book *Post-Mining of Association Rules: Techniques for Effective Knowledge Extraction* [45], the number of association rules generated in fields such as retail, telecom, and insurance is often enormous, with many being of little value. The same problem has been identified in the building sector, as noted in the literature. This significantly increases the time required for domain experts to extract valuable knowledge from the mined association rules. Consequently, post-mining is essential for enhancing the efficiency of identifying valuable association rules.

Combining the cluster and association rule mining approaches, some research has been carried out to investigate building running conditions. Ref. [46] proposes a three-step K-means clustering framework for energy benchmarking using time-series data, improving accuracy by categorizing buildings based on operational similarities. Ref. [47] proposed an effective post-mining workflow with FP-growth to filter and reduce association rules from building operation data, revealing significant patterns and faults. Ref. [19] studied a hybrid data mining-based framework for identifying and interpreting typical electricity load patterns to enhance building energy management and anomaly detection.

Table 1 presents some research that has investigated similar tasks with method significant information. This table shows that most of the articles used single clustering and ARM methods to realize the knowledge discovery mission from the panel data. It is insufficient to determine which algorithm is the best for time-series data analysis in the building investigation field. This research aims to address the existing literature gap concerning the optimal algorithm for time-series data analysis in building energy investigations. While previous studies have primarily employed single clustering and ARM methods for knowledge discovery from panel data, they have not definitively determined the most effective approach for time-series data. This study compares various clustering methods analyzing occupants' continuous electricity demand behavior in building energy management. It contributes by thoroughly exploring various time series clustering algorithms and comparing their efficacy in capturing occupants' power load demand patterns. Additionally, the study introduces novel indicators to assess clustering performance and examines factors influencing disparate clustering outcomes.



**Table 1.** Comparison of different literature.

Author	Time	Data Type	Methods	Unsupervised Algorithms	Purpose
[48]	2015	Panel data	Cluster; regression	k-means	Prediction
[49]	2016	Panel data	Cluster; ARM	k-means; apriori	Knowledge extract
[50]	2018	Longitudinal data	ARM	Gradual pattern mining	Knowledge extract
[51]	2018	Panel data	Cluster	k-means	Knowledge extract
[17]	2018	Cross-section data	Cluster	k-means	Knowledge extract
[30]	2019	Panel data	Cluster	k-means	Knowledge extract
[24]	2019	Panel data	Cluster	k-means	Knowledge extract
[47]	2020	Cross-section data	ARM	FP-growth	Knowledge extract
[46]	2020	Panel data	Cluster	k-means	Knowledge extract
[19]	2021	Panel data	Cluster	k-means	Knowledge extract
[18]	2022	Cross-section data	Cluster; regression	k-means	Prediction
[52]	2023	Panel data	Cluster; regression	k-medoids	Prediction
[20]	2023	Panel data	Cluster	Hierarchical	Knowledge extract
[33]	2023	Panel data	Cluster	k-means	Knowledge extract
[53]	2024	Cross-section data	Cluster	Hierarchical	Knowledge extract
[22]	2024	Panel data	Cluster	k-means; k-shape; DTW, DDTW	Knowledge extract
[54]	2024	Panel data	Cluster	t-SNE	Knowledge extract
[55]	2024	Panel data	Cluster; classification	k-means	Knowledge extract; prediction
[32]	2024	Panel data	Cluster	Deep learning	Knowledge extract

To explore the application prospects of various time-series clustering algorithms in analyzing occupants' continuous electricity demand behavior, this study compares the clustering effects of multiple different algorithms. The contributions of this study are as follows:

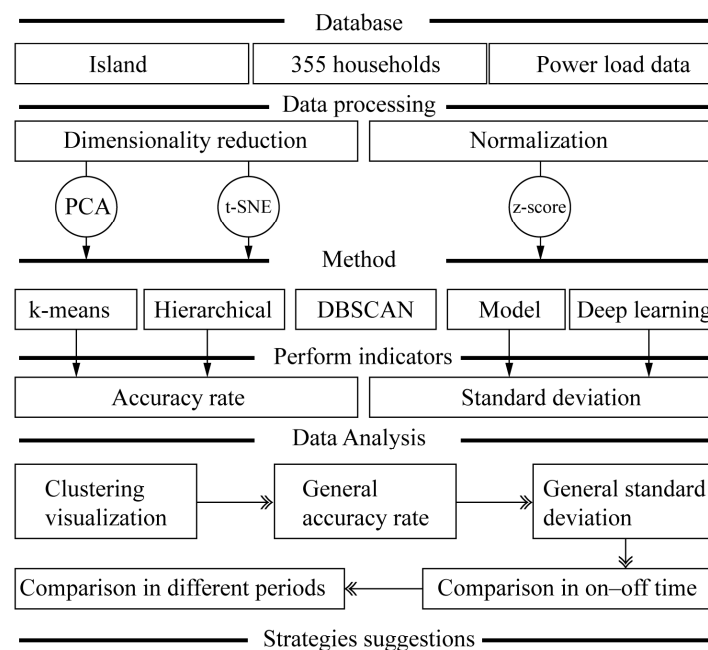
1. Conducted a comprehensive investigation of time series clustering methods for occupants' behavior in the building energy field;
2. Compared the clustering effects of various time series clustering algorithms on occupants' continuous power load demand behavior;
3. Introduced two specific indicators for evaluating the clustering effects of these algorithms;
4. Analyzed the potential reasons for differences in clustering results across different algorithms;
5. Discovered island rural residents' behavior law for energy saving in line with sea island geographic characteristics.

This paper is organized as follows. Section 1 introduces several investigations that have been carried out in recent years, pointing to the shortage of current research. Section 2 indicates the clustered methods used in this paper and interprets the relative computational theory. The third result part (Section 3) analyses the studied results with quantitative performance indicators. The Discussion section (Section 4) focuses on comparing different algorithms and potential reasons for various approaches and typical energy variation phenomenon of rural buildings. In the final part (Section 5), all crucial findings are summarized as conclusions.

## 2. Materials and Methods

### 2.1. Outline

In this paper, Figure 2 presents the basic research steps. First, the database consists of 355 households' time-series energy usage data on a typical island for post-analysis. Secondly, in order to achieve a cluster, the data processing phase introduces two manners of dimensionality reduction and normalization tasks. PCA and t-SNE are used to reduce the data dimension and realize the time-series data visualization. Then, 14 clustering algorithms are performed on the data. These cluster methods can be classified into five categories: k-means, hierarchical, DBSCAN (density-based spatial clustering of applications with noise), machine learning model, and deep learning algorithm.



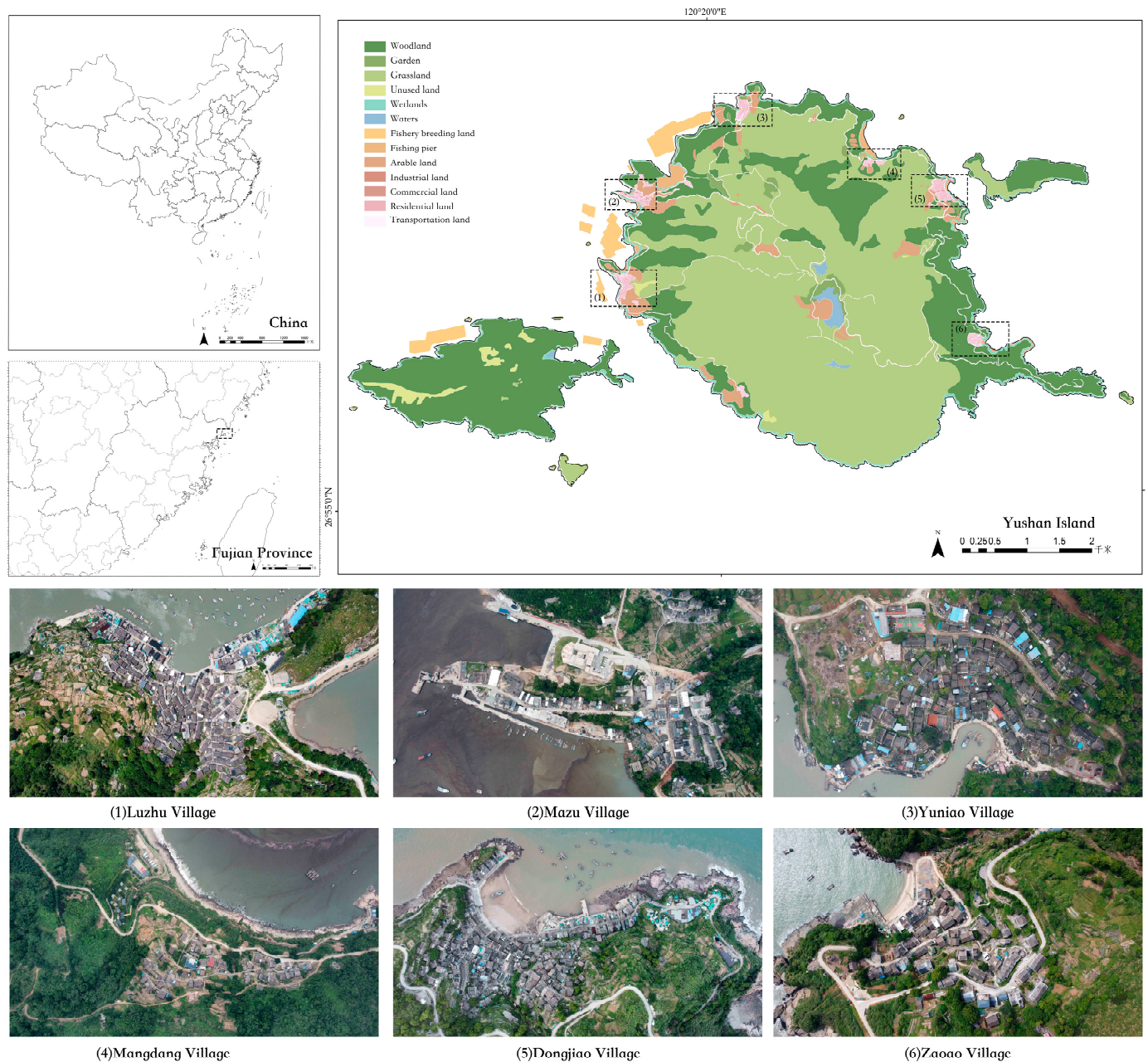
**Figure 2.** Outline of this research.

Macro and micro level comparison is conducted on clustered results using the performance indicators of accuracy rate and standard deviation. These two performance indicators are defined in this paper. In line with the comparison consequences, the best advantage information is the output of cluster algorithms on rural building's energy variation curves. In the data analysis section, the clustered result is presented by general accuracy rate and standard deviation data. On-off time and different periods are the two main measured time for clustering analysis. Finally, some advances for carbon emission reduction could be provided for the determination of government strategies.

### 2.2. Data Sources

#### 2.2.1. Island Site

Due to the precision level of the database, Yushan Island is considered the primary data source. Yushan Island is located in the southeast sea area of Fuding City, Fujian Province. It covers an area of 21.2 square kilometers, has a diameter of approximately 5 km, and a coastline extending 30.12 km. This region features diverse geographical characteristics and falls within the subtropical monsoon climate zone. Summers are hot and humid, with temperatures around 25 °C, while winters are mild and dry, averaging 6 °C. The annual average temperature is 18.8 °C, with the highest recorded temperature at 38 °C and the lowest at minus 1.2 °C. The island is home to six villages, with a total population of 5003 residents. Figure 3 illustrates the position and layout of Yushan Island.



**Figure 3.** Yushan island map. (千米: kilometer).

### 2.2.2. Database Condition

Yushan Island, home to 355 households, has had its electricity usage data provided by the relevant authorities and Yushan town government for each family. This data has been organized into a time-series format from 1 May 2021 to 26 May 2022, resulting in a daily energy consumption curve with 383 data points. These curves, which show variations in energy usage, serve as the basis for clustering analysis. To further understand the energy consumption patterns of each household, 355 questionnaires were also prepared to survey various aspects of power load variation.

Table 2 outlines 14 specific indicators aimed at reflecting the energy consumption habits of each family. However, not all of these indicators may significantly influence energy usage patterns. Therefore, efficient methods are necessary to filter out irrelevant indicators, preparing for the construction of a classifier model. After simplification, the remaining indicators will form the variables of the classifier model.

**Table 2.** Questionnaire options.

Option	Explanation	Option	Explanation
Building Type	Type of resident or public building	Power	Annual electricity consumption
Population	Long-term residents	Power intensity	Level of electricity consumption
Coast	Reside near the seaside or not	Insulation	Presence of insulation material
Job	Primary job type	Equipment	Cooling equipment used in summer
Island	Live on island or not	Age	Average age of householders
Width	Width of rural house	Orientation	Direction of building
Depth	Depth of rural house	Structure	Type of bearing structure

### 2.3. Algorithms

To compare clustering algorithms as much as possible, there are 14 algorithms with five categories, as shown in Table 3. All these manners are performed on the same dataset to achieve the time-series information clustered purpose. Despite the different calculation logic for these approaches, they are evaluated by established performance indicators under the same levels.

**Table 3.** All algorithms for clustering analysis.

Algorithm Type	Calculation Method	Abbreviation
K-MEANS	k-shape	Km_kshape
	Euclidean	Km_Euclidean
	DTW	Km_DTW
	softdtw	Km_softdtw
Hierarchical	Euclidean	Hi_Euclidean
	manhattan	Hi_Manhattan
	DTW	Hi_DTW
Density	DBSCAN	DBSCAN
Model	Hidden Markov model,	HMM
	Auto-regressive model	AR
Deep learning	Recurrent neural network	RNN
	Autoencoder	Auto
	Spectral clustering	SC
	Time-window clustering	TWC

#### 2.3.1. PCA and t-SNE

PCA (principal component analysis) and t-SNE (t-distributed stochastic neighbor embedding) are both techniques used in machine learning and data visualization, particularly for reducing the dimensionality of data to make it easier to explore and visualize.

PCA is a technique for reducing the dimensionality of data by transforming it into a set of orthogonal components that capture the maximum variance. It begins by computing the covariance matrix shown in Equation (1).

$$C = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x})^T \quad (1)$$

where  $x_i$  are the data points and  $\bar{x}$  is the mean vector.  $C$  is the covariance matrix of the data. PCA then finds eigenvectors  $v_j$  and eigenvalues  $\lambda_j$  of  $C$ . The principal components  $u_j$  are selected based on these eigenvalues, representing directions of maximum variance

(Equation (2)). By projecting the data onto these components, PCA reduces its dimensionality while preserving key information for efficient visualization and analysis.

$$u_j = \frac{1}{\sqrt{\lambda_j}} v_j \quad (2)$$

t-SNE, or t-distributed stochastic neighbor embedding, aims to map high-dimensional data points  $x_i$  into a lower-dimensional space  $y_i$  while preserving pairwise similarities as much as possible. It minimizes the Kullback–Leibler divergence between the joint probability distributions of the high-dimensional data  $p_{ij}$  and the low-dimensional embeddings  $q_{ij}$ :

$$C = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}} \quad (3)$$

where  $p_{ij}$  is the similarity between data points  $x_i$  and  $x_j$  in the high-dimensional space, normalized as Equation (4).  $q_{ij}$  represents the similarity between embeddings  $y_i$  and  $y_j$  in the low-dimensional space using the Student's t-distribution:

$$p_{ij} = \frac{\exp(-\|x_i - x_j\|^2 / 2\sigma_i^2)}{\sum_{k \neq l} \exp(-\|x_i - x_j\|^2 / 2\sigma_i^2)} \quad (4)$$

$$q_{ij} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{k \neq l} (1 + \|y_i - y_j\|^2)^{-1}} \quad (5)$$

This process effectively preserves local and global structures of the data, making t-SNE a powerful tool for visualizing complex datasets.

Z-score normalization, also known as standardization, is a method of scaling data so that it has a mean of 0 and a standard deviation of 1. This process is useful in comparing data that have different units or scales, or for preparing data for machine learning algorithms that assume data to be normally distributed.

The formula for Z-score normalization is:

$$Z = (X - \mu) / \sigma \quad (6)$$

where  $Z$  is the z-score,  $X$  is the original value,  $\mu$  is the mean of the dataset, and  $\sigma$  is the standard deviation of the dataset.

### 2.3.2. k-Means

The k-means clustering algorithm is widely used for clustering time-series data, and its effectiveness can be influenced by the distance computational method used. Based on the different distance metrics of Euclidean distance, dynamic time warping (DTW), and soft dynamic time warping (SoftDTW), the k-means algorithm could be sorted into three manners.

Euclidean distance is the most straightforward distance measure and is defined as:

$$d_{euclidean}(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2} \quad (7)$$

where  $x$  and  $y$  are two time-series of length  $n$ .

DTW is a more flexible distance measure that allows for the alignment of the time series with different lengths or nonlinear distortions. The DTW distance between two time-series  $x$  and  $y$  is defined as the minimum cumulative distance required to align them:

$$d_{dtw}(x, y) = \sqrt{\min_{\pi} \sum_{(i,j) \in \pi} (x_i - y_j)^2} \quad (8)$$



where  $\pi$  is a warping path that defines a mapping between indices of  $x$  and  $y$ . The DTW distance is computed using dynamic programming. Let  $D(i, j)$  be the cumulative distance up to point  $(i, j)$ :

$$D(i, j) = (x_i - y_j)^2 + \min(D(i-1, j), D(i, j-1), D(i-1, j-1)) \quad (9)$$

The final DTW distance is  $D(n, m)$  where  $n$  and  $m$  are the lengths of  $x$  and  $y$ , respectively.

SoftDTW is a differentiable version of DTW that provides a smoother cost function, which is useful for optimization. It replaces the minimum operation in DTW with a soft minimum, which can be defined using a smoothing parameter  $\gamma$ . The SoftDTW distance is computed similarly to DTW, but using the soft minimum:

$$S(i, j) = (x_i - y_j)^2 + \text{softmin}_{\gamma}(S(i-1, j), S(i, j-1), S(i-1, j-1)) \quad (10)$$

The parameter  $\gamma$  controls the smoothness, with larger values making the function smoother.

K-Shape is a time-series clustering algorithm that focuses on aligning and clustering time-series data based on their shapes. It uses a shape-based distance measure, which is invariant to scaling and shifting, making it particularly suitable for time-series data. Fundamentally speaking, it also belongs to the k-means method.

The shape-based distance measure in K-Shape is defined using the cross-correlation between z-normalized time series. For two z-normalized time-series  $X$  and  $Y$ , the shape-based distance is defined as:

$$SBD(\vec{x}, \vec{y}) = 1 - \max_w \left( CC_W(\vec{x}, \vec{y}) / \sqrt{R_0(\vec{x} \cdot \vec{x}) \cdot R_0(\vec{y} \cdot \vec{y})} \right) \quad (11)$$

where  $\tau$  is the lag and  $(i + \tau) \bmod n$  ensures circularity.

### 2.3.3. Hierarchical Clustering Algorithm

Hierarchical clustering is a method for creating a hierarchy of clusters for time-series data. Unlike partitioning methods like k-means, hierarchical clustering does not require the number of clusters to be specified in advance. Instead, it builds a dendrogram, a tree-like structure that represents the nested grouping of the data based on a chosen distance metric. Similarly, in terms of distance calculation, Euclidean distance, DTW, and Manhattan could also be used.

Manhattan distance is defined as the sum of the absolute differences between corresponding points of the two time series:

$$d_{\text{manhattan}}(X, Y) = \sum_{i=1}^n |X_i - Y_i| \quad (12)$$

### 2.3.4. DBSCAN

DBSCAN (density-based spatial clustering of applications with noise) is a popular clustering algorithm that is particularly effective at identifying clusters of varying shapes and densities in time-series data. Unlike traditional clustering methods like k-means, DBSCAN does not require the number of clusters to be specified beforehand and can automatically identify outliers as noise. The distance metrics are also the same as the aforementioned approaches. The working theory is as follows:

For a given point  $p$ , its  $\epsilon$ -neighborhood  $N_{\epsilon}(p)$  consists of all points within a distance  $\epsilon$  from  $p$ , as shown in Equation (13):

$$N_{\epsilon}(p) = \{q | \text{distance}(p, q) \leq \epsilon\} \quad (13)$$

where distance is typically a metric like Euclidean distance. Point  $p$  is considered a core point if the number of points within its  $\varepsilon$ -neighborhood is at least a given threshold  $\text{minPts}$ . Therefore,  $p$  is a core point if it meets the limitation condition Equation (14):

$$|N_\varepsilon(P)| \geq \text{minPts} \quad (14)$$

where  $\text{minPts}$  is a parameter that determines the minimum number of points required to form a dense region. Point  $q$  is density-reachable from a point  $p$  if there exists a chain of core points where each core point is within the  $\varepsilon$ -neighborhood of the previous one, and  $q$  is in the  $\varepsilon$ -neighborhood of the last core point in the chain. DBSCAN starts by identifying all core points based on the  $\varepsilon$ -neighborhood and the  $\text{minPts}$  threshold. It then forms clusters by connecting core points that are density-reachable from each other.

Points that are not core points but fall within the  $\varepsilon$ -neighborhood of core points are added to the cluster of the core point. Finally, points that are not reachable from any core point are labeled as noise.

### 2.3.5. Model

Hidden Markov models (HMMs) are probabilistic models that assume the system being modeled is a Markov process with unobserved (hidden) states. HMMs are particularly useful for time-series data as they can model the temporal dependencies and underlying structure.

Auto-regressive (AR) models are linear models that predict future values of a time series based on its own past values. An AR model of order  $p$  ( $\text{AR}(p)$ ) is given by:

$$X_t = c + \sum_{i=1}^p \phi_i X_{t-i} + \epsilon_t \quad (15)$$

where  $X_t$  is the value of the time series at time;  $t$  is a constant;  $\phi_i$  are the parameters of the model;  $\epsilon_t$  is white noise (random error term).

### 2.3.6. Deep Learning

Recurrent neural networks (RNNs) are a class of neural networks that are particularly well-suited for handling sequential data such as time series. RNNs have a unique architecture that allows them to maintain information about previous inputs in their internal state, which makes them effective for capturing temporal dependencies. RNNs process sequences of data one step at a time, maintaining an internal state that captures information about previous steps. This internal state allows the RNN to exhibit temporal dynamic behavior, which is crucial for time-series analysis.

Hidden state update is calculated as follows:

$$h_t = \sigma(W_h h_{t-1} + W_x X_t + b_h) \quad (16)$$

where  $W_h$  is the weight matrix for the hidden state;  $W_x$  is the weight matrix for the input;  $b_h$  is the bias term, and  $\sigma$  is the activation function. The output calculation is

$$y_t = W_y h_t + b_y \quad (17)$$

where  $W_y$  is the weight matrix for the output and  $b_y$  is the bias term.

Autoencoders are a type of neural network used to learn efficient codings of input data. They are particularly useful for dimensionality reduction and feature extraction, making them well-suited for clustering time-series data. An autoencoder consists of two main parts: an encoder that compresses the input into a latent space representation and a decoder that reconstructs the input from this representation.

In addition, spectral clustering is another technique that leverages the eigenvalues (spectrum) of a similarity matrix derived from the data to perform dimensionality reduction before clustering in fewer dimensions. It is particularly effective for identifying clusters in data that is not well-separated in a traditional Euclidean space, including time-series data.

Apart from the above algorithms, time-window clustering is a special approach that divides time-series data into smaller, more manageable segments (time windows) and then performs clustering on these segments. This method is particularly useful for identifying patterns or behaviors that vary over time within the same time series. The steps for time-window clustering typically include segmenting the data, extracting features from each segment, and applying a clustering algorithm to these features.

## 2.4. Perform Indicators

### 2.4.1. Accuracy Rate

First of all, the accuracy rate is the most significant purpose of a cluster. Consequently, this investigation first takes the manual recognized energy variation pattern as the basic standard. This typically involves observing, analyzing, or interpreting data manually to recognize specific trends or variations in energy over time or across different conditions. In essence, the pattern was detected through careful examination and analysis by multiple researchers. These recognition results are defined as the correct energy variation pattern. All clustered results by various algorithms are compared with this defined criterion. Parameter  $A_{acc}$  is defined as follows in Equation (18) to measure the clustered precision.

$$A_{acc} = (\text{the number of correct cluster sample}) / (\text{the number of all samples}) \quad (18)$$

In this case, the larger  $A_{acc}$  represents a greater cluster algorithm of higher accuracy.

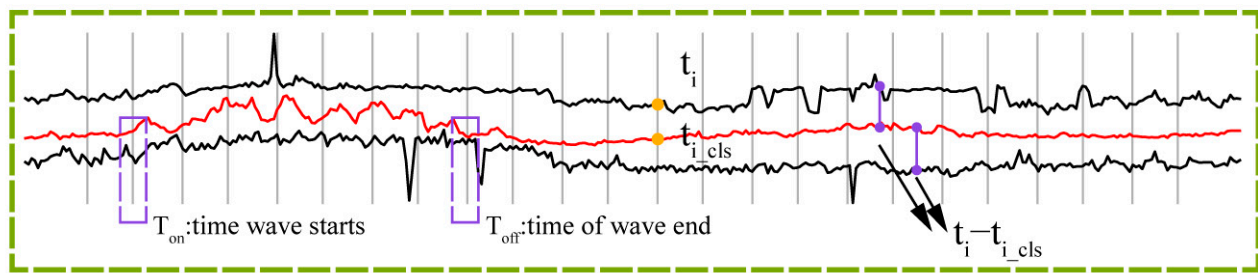
### 2.4.2. Standard Deviation

Traditionally, there are currently many methods evaluating clustering results, such as the sum of squared errors (SSE), the silhouette coefficient (SC), the Calinski Harabasz index (CH), the Davies Bouldin index (DB), and the Dunn index (DI). However, these metrics primarily search for the  $k$  value instead of the final clustering performance. Therefore, it is indispensable to define a new index to measure the energy curve features. Ideally, the clustering curve should be perfectly parallel to the energy curves. That is to say, the difference between the cluster and basic data are the same at each time. Hence, the standard deviation of the difference between the cluster and the sample values (SD) is defined as Equation (12) to measure the clustering result.

$$SD = \sqrt{\frac{\sum_{i=1}^n [(t_i - t_{i-cls}) - (\bar{t} - t_{cls})]^2}{n}} \quad (19)$$

where  $t$  is the energy value for sample data,  $i$  refers to the time,  $t_{cls}$  means the energy value of the cluster center, and  $n$  represents the total number of time points. According to this calculation method, and the smaller SD value represents the better clustering performance.

Beyond the SD being used for assessing overall clustered performance, it also needs to observe and compare information in detail. Thus, the beginning time of the crest and trough is also recorded to describe the clustered energy profile. It could also be called as the energy variation starts to change time.  $T_{on}$  means the time the wave starts and  $T_{off}$  refers to the time of the wave end, as shown in Figure 4.



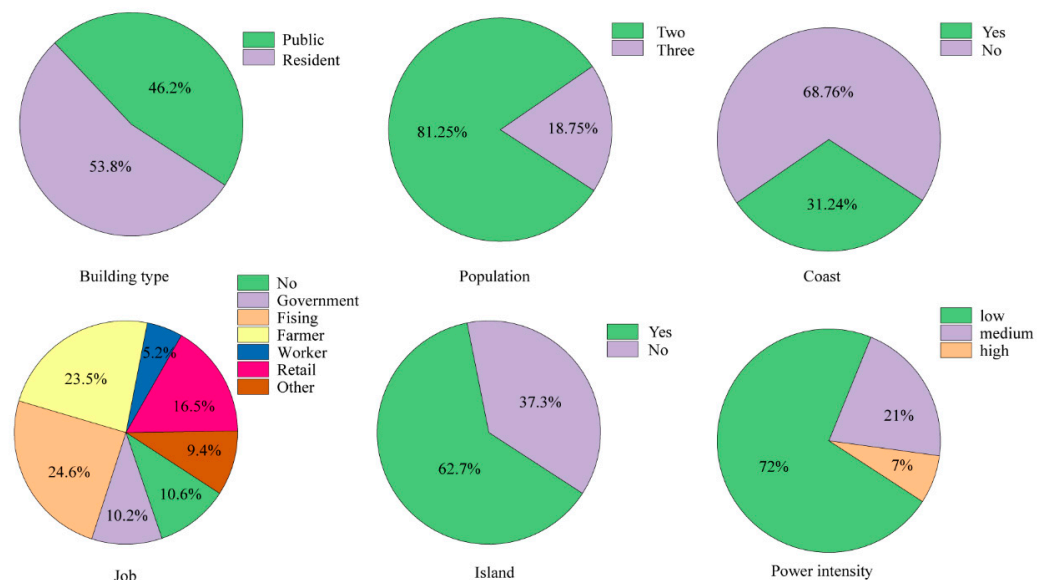
**Figure 4.** Performance indicator of the cluster calculation method.

### 3. Results

#### 3.1. Questionnaire Results

The questionnaire aims to gather data on building factors to analyze how each correlates with carbon emissions from buildings. Researchers investigated a total of 355 residents face-to-face over two months. As we surveyed these questions in person, the response rate was 100%. Meanwhile, the validation efficiency was also 100%. In this case, 355 records were used to analyze the clustering performance for different algorithms and study the relationship between building attributes and clustered results.

Figure 5 presents the statistical results of the questionnaire. For the investigated samples, the number of residences was almost equal to that of the public-type. Most families consisted of two people because the young people lived out of island to make money. For islanders, a large number of people work in fishing and agriculture, while the least number of people work in secondary industries. In term of energy usage structure, as there are many elders, the house energy consumption mainly remains at low level. In addition, more families usually live on the island in the investigated samples.



**Figure 5.** Questionnaire statistical results.

#### 3.2. Clustered Accuracy Rate

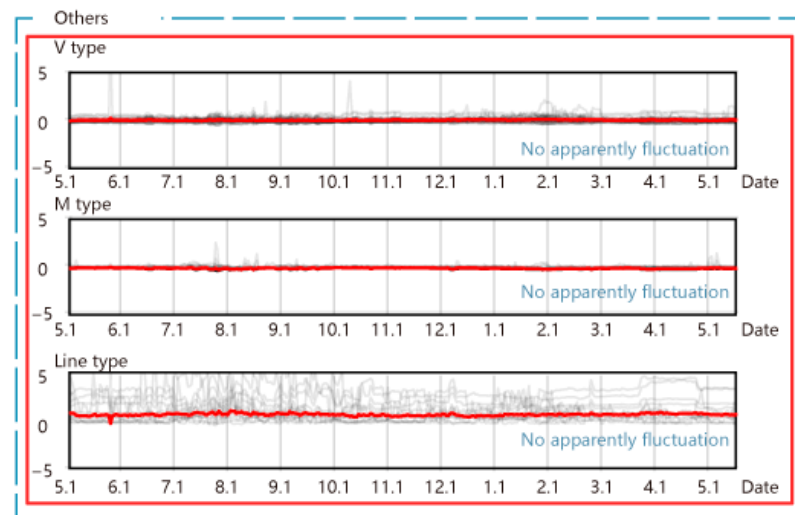
After clustering for all energy data, the overall shape performance and accurate rate are shown in Table 4. All algorithms could cluster energy data into three types of M, V, and Line. The M type illustrates two significant wave peaks, and the V pattern shows some incisive waves in a short period. The line-type curve has no distinct variation throughout the whole year.

**Table 4.** Accuracy rate for different algorithms.

Algorithms	M		V		LINE		Total	
	Y	N	Y	N	Y	N	Y	N
Km_kshape	66.67%	33.33%	81.25%	18.75%	41.18%	58.82%	67.05%	32.95%
Km_Euclidean	69.23%	30.77%	73.38%	26.62%	60.10%	39.90%	65.09%	34.91%
Km_DTW	84.62%	15.38%	53.13%	46.88%	35.29%	64.71%	63.64%	36.36%
Km_softdtw	84.62%	15.38%	53.13%	46.88%	35.29%	64.71%	63.64%	36.36%
Hi_Euclidean	97.44%	2.56%	0.00%	100.00%	58.82%	41.18%	54.55%	45.45%
Hi_Manhattan	89.74%	10.26%	0.00%	100.00%	41.18%	58.82%	47.73%	52.27%
Hi_DTW	97.44%	2.56%	0.00%	100.00%	5.88%	94.12%	44.32%	55.68%
DBSCAN	76.92%	23.08%	12.50%	87.50%	17.65%	82.35%	42.05%	57.95%

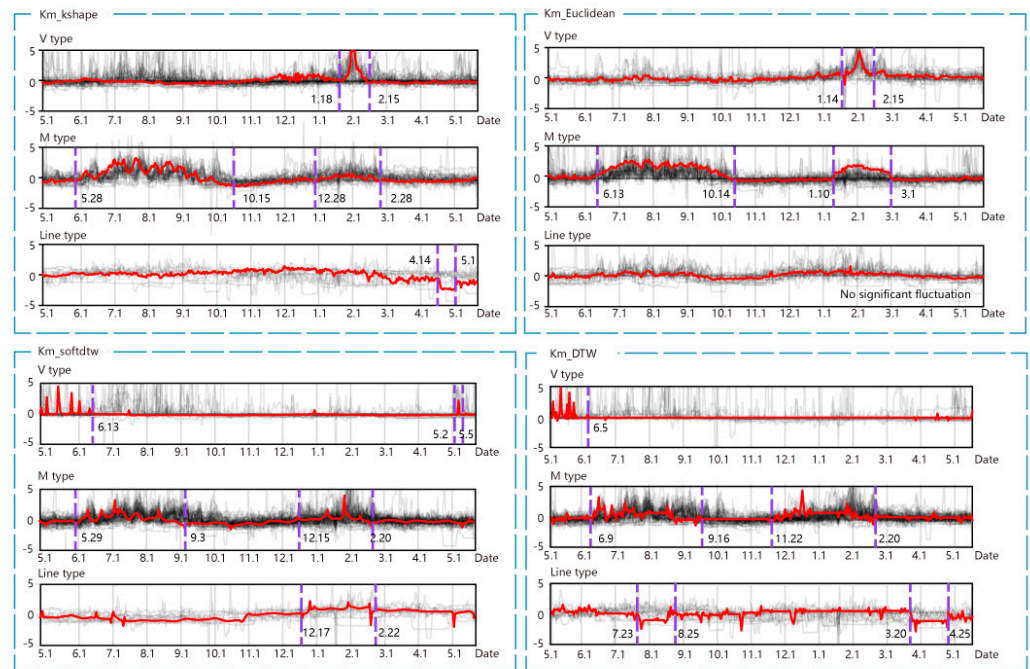
(Km\_kshape: k-shape algorithm; km\_Euclidean: k-means algorithm with Euclidean distance calculation; Km\_DTW: k-means algorithm with dynamic time warping distance calculation; Km\_softdtw: k-means algorithm with soft dynamic time warping; Hi\_Euclidean: hierarchical clustering algorithm with Euclidean distance calculation; Hi\_Manhattan: hierarchical clustering algorithm with Manhattan distance calculation; Hi\_DTW: hierarchical clustering algorithm with dynamic time warping distance calculation; DBSCAN: density-based spatial clustering of applications with noise).

The result shows that in the algorithms of the model, deep learning fails in grouping energy variation data (Figure 6). It can be seen that each energy consumption curve, no matter in what form, is clustered into the line type. No variation features are identified via intelligent algorithms. Therefore, these complicated methods are not suitable for time-series building power loads cluster analysis. This is because a highly complicated machine learning approach always requires a long-time parameter tuning procedure, which has bad achievement under default conditions.

**Figure 6.** Clustered result of algorithms with model, deep learning.

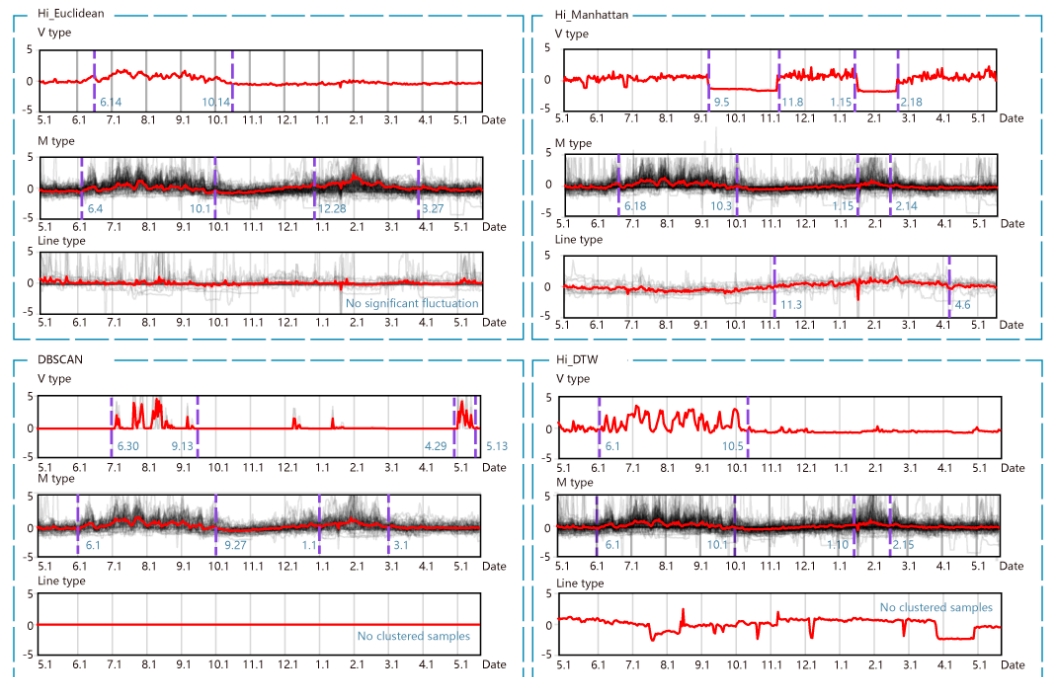
Secondly, in general, each k-means manner indicates an imitative effect, while hierarchy algorithm performances vary greatly under different distance calculation methods (Figure 7). Combining with the accuracy rate, the k-shape presents the highest total accuracy rate, especially for the V pattern identification aspect. For k-means with DTW and softdtw distance computation approach, overall, Aacc is slightly lower than k-shape, but this rate is higher in terms of M-type recognition. Meanwhile, adopting the Euclidean distance calculation method of k-means has better cluster performance. For example, the V shape and line curve are identified correctly only with 73.38% and 60.10%, respectively.





**Figure 7.** Clustered results for different k-means algorithms.

In terms of hierarchy clustered manners, each algorithm could cluster the M type under a high accuracy rate (Figure 8). In contrast, in terms of V pattern recognition, these algorithm presents the remarkably poor correct situation as shown in the table above, with a 0.00% accuracy rate. Moreover, in terms of DBSCAN, the M type indicates a good recognition phenomenon; however, the V and Line type curves are clustered badly and have a low rate. In consequence, k-means series algorithms perform well in clustering time-series energy variation curves.

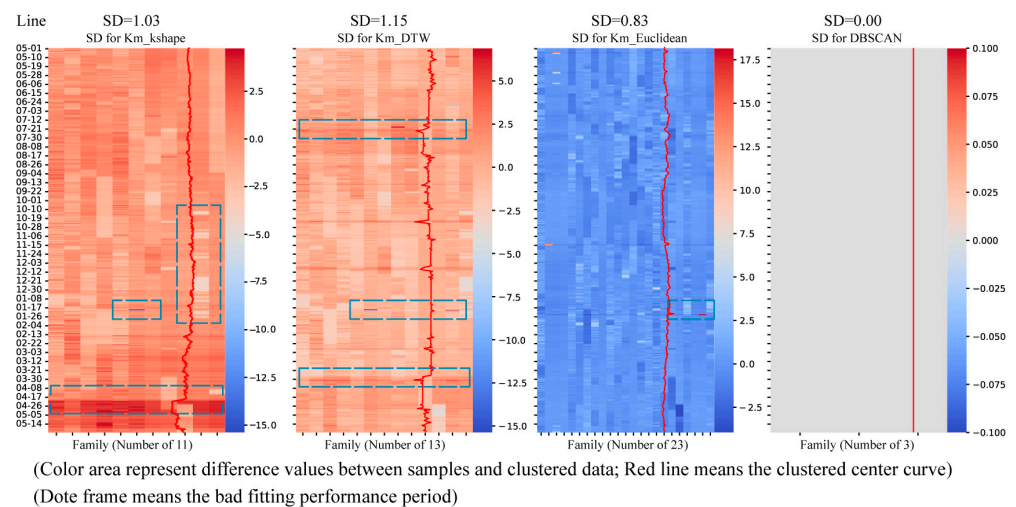


**Figure 8.** Clustered performance of the hierarchical method and DBSCAN.

### 3.3. Clustered SD Values Comparison

SD values could measure the shape-fitting extent between samples and clustered centers. The smaller values correspond to better grouping performance. According to the clustered accuracy rate, four Km algorithms and the DBSCAN method could successfully identify time-series curves. Considering the Km\_DTW and Km\_softdtw have the same results, the remaining four algorithms were analyzed via the SD indicator.

Figure 9 presents the difference value between the clustered center and sample data for the above four algorithms. First of all, DBSCAN's SD values are the lowest among all the methods. However, the line and V types only contain three and four sample data curves, respectively (Figure 9). That means the clustering generalization ability is weak because most of the data are classified into the M type; for example, the corresponding SD value of M reached 1.88.

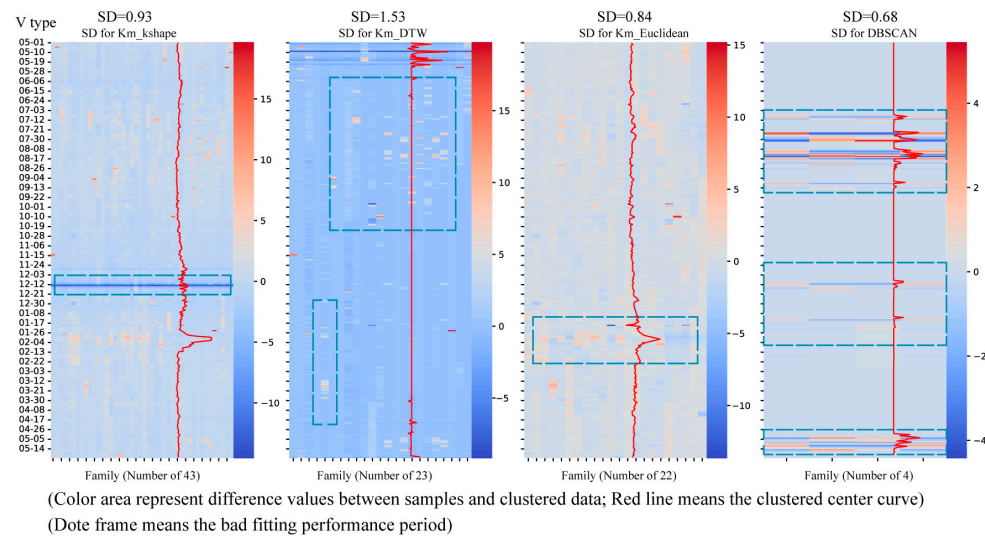


**Figure 9.** SD performance for line type.

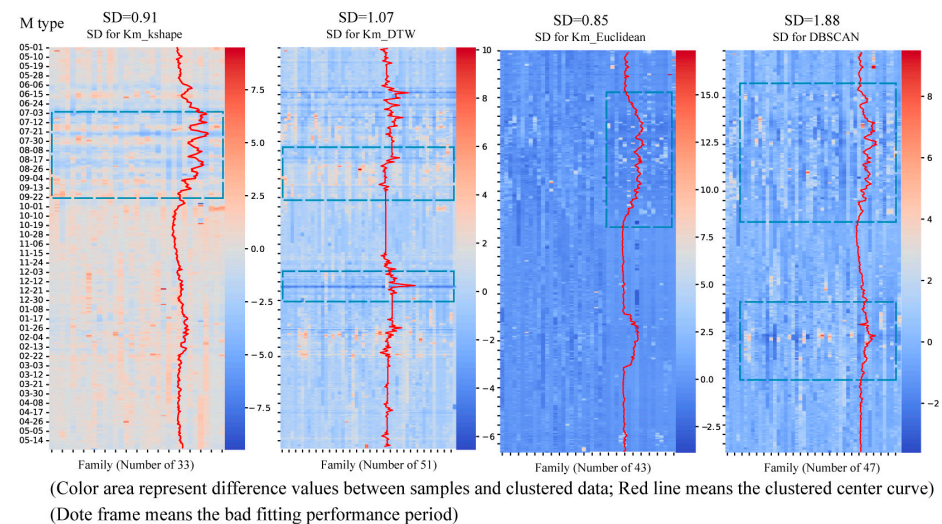
Secondly, Km\_kshape and Km\_DTW have the best identification result for M type comparing other patterns. Meanwhile, Km\_Euclidean indicates a good fitting effect in terms of all three patterns, illustrating that the clustered center curve could precisely reflect these time series information features. Thirdly, regarding V type, strike wave matching can not be achieved by Km\_DTW successfully (Figure 10). Meanwhile, for Km\_kshape, better fitting results were shown in the initial phase of the wave crest. In general, Km\_kshape had a better-clustered performance than Km\_DTW for the V pattern, which was shown by a lower SD value.

In terms of the M type, despite the Km\_kshape being similar to Km\_DTW for the fitting effect with an identical SD, the matching presentation of Km\_DTW was worse than Km\_kshape during the wave period, as shown in Figure 11 (green frame). It can be observed that the differences between the clustered centers and samples varied significantly during the crest of the wave. Therefore, Km\_DTW is more suitable for recognizing the steady time-series data compared to fluctuating data. Lastly, for the line type, Km\_kshape failed to identify the abnormal data after 7 May. Contrarily, Km\_DTW could successfully fit these exceptional data into the line type. Thus, Km\_DTW has better performance in terms of processing unusual data than Km\_kshape, which presents a lower SD value.

Overall, the DBSCAN algorithm is not conducive to promote because of its bad generalization capability. Km\_Euclidean presents the best clustering effect and fitting performance. Other algorithms have their advantages, respectively. Km\_kshape is appropriate for normal time-series energy data; however, it is weak concerning abnormal samples. Moreover, Km\_DTW could dispose of steady and abnormal data. However, with info varying considerably in a short time, it can not achieve a great fitting performance.



**Figure 10.** The SD performance for the V type.



**Figure 11.** The SD performance for the M type.

## 4. Discussion

### 4.1. Clustering Detail Comparison for Different Algorithms

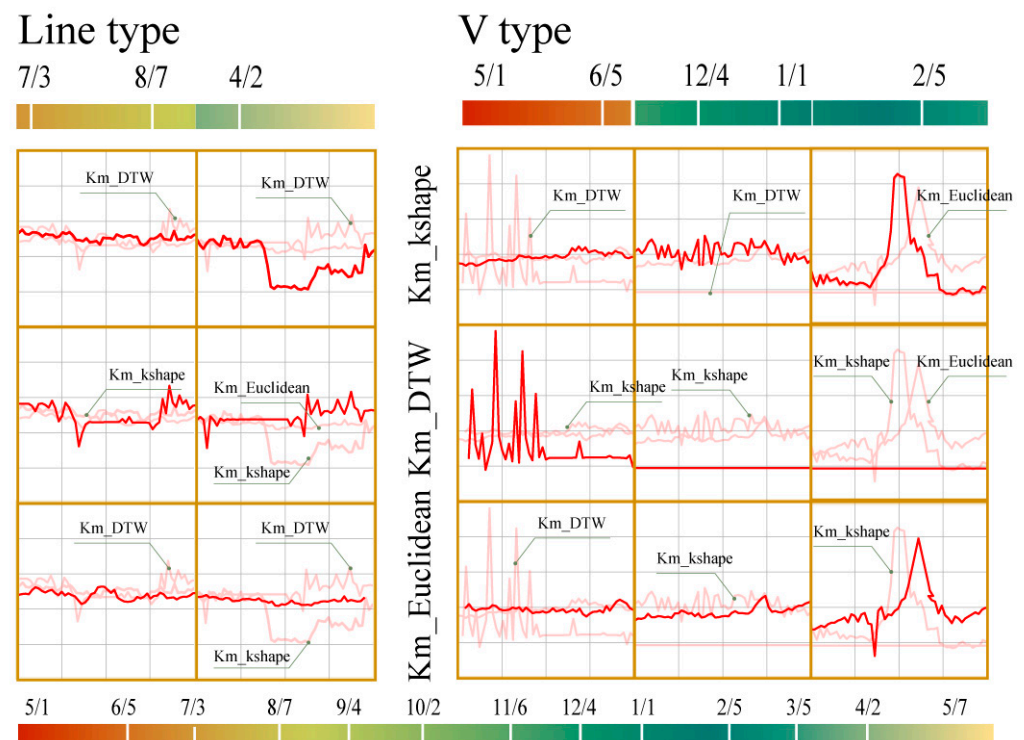
Beyond general identification accuracy evaluation, the three algorithms that successfully identify time-series curves are required to be compared in detail. Figure 12 presents a detailed comparison of the V and line manners. The gradient block means the timeline for power load variation. It can be clearly seen the clustered center results in differences for various clustered types under the same period.

In terms of the V type, Km\_DTW significantly misidentifies the smooth as the intermittent wave from 1 May to 5 June. Meanwhile, around 5 February, it also classified the single peak wave into a steady line inaccurately. This phenomenon indicates that the Km\_DTW is poor at identifying unimodal electricity curves and is liable to mistakenly classify the less volatile curve as a V-shaped. On the other hand, Km\_kshape and Km\_Euclidean could successfully recognize the V type curve features. Both of them could regard the less volatile curve as the line shape, ignoring the disturbing fluctuation points.

With respect to the M type, Km\_DTW could also classify the soft waveform from 5 June to 4 September, like Km\_kshape and Km\_Euclidean (Figure 13). However, the clustered center performance of Km\_DTW displays more intermittently than the other two algorithms. That is to say, it magnifies some wave peaks and reduces meaningful valley



curve information. In this matter, the continuous wave trend is clustered into intermittent figures such as the electricity changing between 4 December and 5 March. Similarly, for line type investigation, Km\_DTW also illustrates analogous grouping performance, which changes the soft curve into intermittent lines. This consequence is in accordance with the latest research of [22]. It found that the Km\_DTW has a better effect focusing on intermittent building energy data. While this research adds and proves that the clustering effect using the DTW distance calculation method with k-means on continuous energy data is not good.

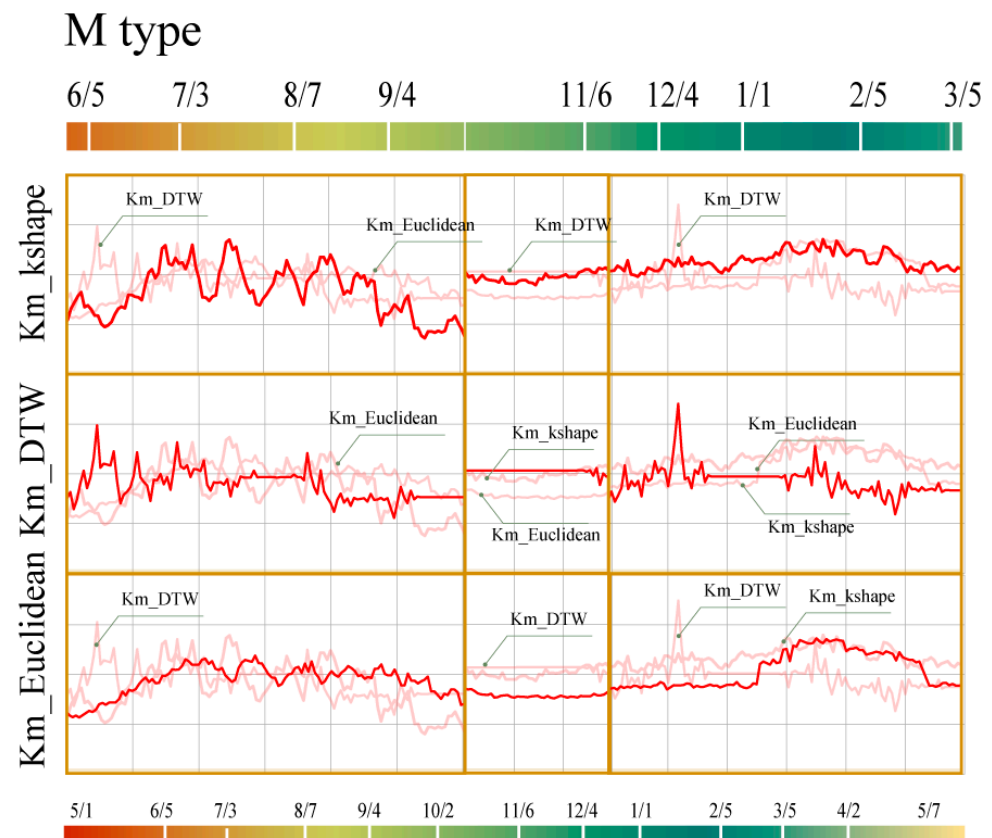


**Figure 12.** Comparison for detail clustered patterns of the V and Line types.

For the other two algorithms, the M type performance shows that Km\_shape has a better identification presentation than Km\_Euclidean with respect to the place of the wave start. Km\_Euclidean distinguished and ignored part wave rising information at the beginning of the wave. This is because the Euclidean distance computational approach abandons shape characteristics and only depends on numeric data calculation. The Km\_kshape disposes of this data smoothly from the remaining basic pattern information. This SBD calculation method could capture the pattern form characteristics without considering the data itself. Hence, Km\_kshape does better in the clustered operation for the M type than Km\_Euclidean. Taking into consideration the line type, an apparent distinction between these two algorithms concentrates on abnormal data conduct. The database used in this research contains some outliers after 2 April. Facing this issue, the k-shape algorithm captures this special curve shape. However, this accurate operation breaks the clustered line type, influencing power load law identification instead. Contrarily, Km\_Euclidean closes the outlier trend gap and classifies it into line type that successfully picks up the curve feature. Therefore, it can be observed that Km\_Euclidean is better than Km\_kshape in terms of generalization performance.

In a word, all these three algorithms could achieve the time-series energy consumption data clustering mission. Therein, Km\_DTW is applicable to intermittent curves instead of continuous data. It recognizes distortion in continuous information that changes dramatically. Regarding the other two algorithms, Km\_kshape and Km\_Euclidean methods could group the power load curve with high precision. This is consistent with the study

by [23]. It compared k-means, bisecting k-means, and the Gaussian mixture model algorithms and found that k-means was the most appropriate to investigate building electricity load patterns based on a dataset containing 191 residential and 164 non-residential buildings. This investigation further extends this achievement, providing multiple algorithm advantages and disadvantages by comparing. Therein, Km\_Euclidean performs with great expression for building electricity load curves with several abnormal data; moreover, for an accurate database without outliers, the Km\_kshape could cluster the pattern features more efficiently.



**Figure 13.** Comparison for detail clustered patterns of the M type.

In addition, in terms of intelligent algorithms, deep learning algorithms often struggle with time series data clustering by default parameters may lack adequate feature extraction tailored for temporal patterns, fail to account for temporal dependencies without specialized architectures (like RNNs or LSTMs), and may not include necessary pre-processing steps (e.g., normalization, smoothing). Additionally, the need for labeled data can be a limitation in clustering tasks, and integration challenges between deep learning features and clustering algorithms can affect performance. Furthermore, the difficulty in tuning numerous hyperparameters can prevent optimal clustering results. Effective clustering typically requires customized model architectures, proper data pre-processing, and careful alignment with clustering methods.

For models, the hidden Markov models (HMMs) and auto-regressive models (AR) may not achieve high performance in time series clustering due to their inherent limitations. HMMs assume a simple state-transition structure that may not capture complex or non-Markovian data patterns, and their performance can be hampered by the need for pre-defined states and computational complexity. AR models, on the other hand, assume linear relationships and fixed memory of past values, which can be inadequate for capturing nonlinear patterns or long-range dependencies in time series data, and they often require the data to be stationary. These limitations can lead to suboptimal clustering results.



#### 4.2. Yushan Island Energy Usage Patterns

Table 5 presents the household type on Yushan Island. After clustering for all energy usage curves, three types of V, M, and line are generated. Combined with user conditions, some energy usage phenomena and characteristics were studied as follows.

**Table 5.** User types on the island.

Building Type		Residence	
Resident	Often on island	Non-fisherman	Middle age
		Fisherman	Aged
	Not on island frequently	-	-
		Often use	-
Public	Government	Non-use frequently	
	Street lamp		
	Processing industry		

The V-shaped energy consumption curve mainly reflects public buildings and some temporary residential houses on the island. Staff working part-time causes sharp peaks in energy usage. Public construction of seafood processing buildings shows similar patterns. Concentrating seafood production in October can reduce sharp spikes and energy wastage. The V-shaped power load pattern corresponds to five user types in public and residential buildings. Renewable energy solutions like solar panels or micro-wind turbines are beneficial for infrequently used public buildings and vacant residential properties. Insulation methods for older individuals and passive cooling strategies for younger residents can reduce winter heat demand and summer air conditioner spikes, respectively.

The M-type energy consumption pattern on the island suits frequent residents, with summer peaks due to air conditioning and winter peaks from high-intensity activities like the spring festival. Residents show notable fluctuations from July to October and December to March. Fishermen also follow this pattern but have lower consumption from August to December when at sea. For energy efficiency, common residences benefit from PV systems and passive design strategies. The government should install solar PV panels or micro-wind generators in fishermen's vacant homes during expeditions to generate clean electricity efficiently.

The line-shaped energy consumption curve reflects different architectural characteristics. Households with seniors (over 70) show minimal variations in usage due to frugal habits and lack of cooling needs. Large public buildings and street lamps display steady patterns, the former due to continuous operation and the latter indicating energy wastage. Solutions to reduce carbon emissions include improving insulation and ventilation for senior residents, turning off certain machines in public buildings, and adjusting street lamp usage to match daylight length.

#### 4.3. Future Limitation

This paper compares several algorithms for clustering of building energy consumption curves. The optimal method has been found based on some performance indicators. However, some issues should be paid for further investigation. Firstly, this research found some advantages and disadvantages for each manner; however, the reasons behind this were not investigated in depth. Every algorithm is constituted by complicated mathematical logic and equations. The identification accuracy could be improved by adapting internal parameters. In this case, future relative research should focus on the mathematical theory of these algorithms, leading to classification mistakes.

Secondly, unsupervised data mining methods generally contain cluster and association rule mining tasks. Currently, research focuses mainly on user group databases. However, several more complex data mining missions require a synthetic working process that integrates the advantages of cluster and association rule mining approaches. Hence, how to combine these two missions into an entire workflow is another valuable issue for further study.

Lastly, with the development of artificial intelligence, whether deep learning algorithms have more accurate performance after adjusting parameters. These enhancements in performance could be attributed to fine-tuning hyperparameters, optimizing neural network architectures, and leveraging large datasets, leading to more precise and reliable AI models across various applications. Future related studies could assess these performances using different parameter-adjusting methods.

## 5. Conclusions

This research aims to study the optimal cluster algorithm for building time series energy consumption. There are 17 clustered methods that are compared and evaluated via multiple performance indicators. Yushan island in Fujian province of China is selected as the studied object for investigation. More than 100 household electricity data are recorded yearly, and the basic dataset is constructed. All clustering methods are performed on this same dataset to compare classification performance. Finally, some significant conclusions are found as follows:

- When using data mining to cluster occupants' energy-using behavior patterns, K-means, which relies on Euclidean distance and k-shape, has traditionally been the primary algorithm. These two methods are the primary choice for similar tasks.
- Km\_DTW is applicable to intermittent curves instead of continuous data. Km\_Euclidean performs great expression for building electricity load curves with several abnormal data; moreover, for an accurate database without outliers, Km\_kshape could cluster the pattern features more efficiently.
- Hierarchy and DBSCAN clustered manners fail to group the time-series energy consumption curves, especially for unimodal curve type. Deep learning algorithms also can not cluster time-series building electricity usage data under default parameters in high precision.
- When clustering using four different distance algorithms, the difference in the time of curve condition changes in the energy pattern ranges from 0 days at the minimum to 14 days at the maximum. This indicates that different algorithms have a similar ability to identify the important time of condition variation.
- Accuracy rate and standard deviation introduced in this study serve as evaluation methods for clustering analysis of continuous electricity demand. These metrics effectively describe the characteristics of curve fitting extent within the clustering results.
- In terms of island carbon emission reduction, during fishing expeditions, it is crucial to utilize fisherman households when vacant. For non-fisherman residents, prioritizing heat insulation during winter is essential for the elderly, while passive design strategies are better suited for middle-aged accommodations. Renewable techniques can be applied to infrequently used public buildings like village committees, presenting an opportunity for significant energy efficiency improvements.

**Author Contributions:** Conceptualization, H.Z. and X.L.; methodology, S.Z.; software, X.W.; validation, J.Y. and R.W.; formal analysis, J.W.; investigation, Z.L.; writing—review and editing, X.L. All authors have read and agreed to the published version of the manuscript.

**Funding:** The investigation reported in this paper is supported by the National Natural Science Foundation No. 52078265; the Post-doctoral Science Foundation No. 043291010; the Ministry of Housing and Urban-Rural Development Project No. R20220430; Opening Funds of the State Key Laboratory of Building Safety and Built Environment and the National Engineering Research Center of Building

Technology and thanks for E-Surfing Digital Life Technology Co., Ltd.; Fuding city, Yushan Town government project No. 20222001104 support for database provision.

**Data Availability Statement:** The data presented in this study are available on request from the corresponding author on reasonable request. The data are not publicly available due to privacy policies.

**Conflicts of Interest:** The authors declare no conflicts of interest. The authors declare that this study received funding from E-Surfing Digital Life Technology Co., Ltd. The funder was not involved in the study design, collection, analysis, interpretation of data, the writing of this article or the decision to submit it for publication.

## References

- Chenari, B.; Carrilho, J.D.; Silva, M.G.D. Towards sustainable, energy-efficient and healthy ventilation strategies in buildings: A review. *Renew. Sustain. Energy Rev.* **2016**, *59*, 1426–1447. [\[CrossRef\]](#)
- IEA. *Energy Technology Perspectives Scenarios*; International Energy Agency (IEA): Paris, France, 2012.
- IRENA. *Renewable Capacity Statistics 2021*; International Renewable Energy Agency: Abu Dhabi, United Arab Emirates, 2021.
- IRENA. *Renewable Energy Statistics 2022*; The International Renewable: Abu Dhabi, United Arab Emirates, 2022.
- IRENA. *World Energy Transitions Outlook: 1.5C Pathway*; International Renewable Energy Agency: Abu Dhabi, United Arab Emirates, 2021.
- Allcott, H.; Mullainathan, S. Behavior and Energy Policy. *Science* **2020**, *327*, 1204–1205. [\[CrossRef\]](#) [\[PubMed\]](#)
- Fan, C.; Xiao, F.; Li, Z.; Wang, J. Unsupervised data analytics in mining big building operational data for energy efficiency enhancement: A review. *Energy Build.* **2018**, *159*, 296–308. [\[CrossRef\]](#)
- Miller, C.; Nagy, Z.; Schlueter, A. A review of unsupervised statistical learning and visual analytics techniques applied to performance analysis of non-residential buildings. *Renew. Sustain. Energy Rev.* **2018**, *81*, 1365–1377. [\[CrossRef\]](#)
- Zhao, Y.; Zhang, C.; Zhang, Y.; Wang, Z.; Li, J. A review of data mining technologies in building energy systems: Load prediction, pattern identification, fault detection and diagnosis. *Energy Built Environ.* **2020**, *1*, 149–164. [\[CrossRef\]](#)
- Li, K.; Ma, Z.; Robinson, D.; Ma, J. Identification of typical building daily electricity usage profiles using Gaussian mixture model-based clustering and hierarchical clustering. *Appl. Energy* **2018**, *231*, 331–342. [\[CrossRef\]](#)
- Rajabi, A.; Eskandari, M.; Ghadi, M.J.; Li, L.; Zhang, J.; Siano, P. A comparative study of clustering techniques for electrical load pattern segmentation. *Renew. Sustain. Energy Rev.* **2020**, *120*, 109628. [\[CrossRef\]](#)
- Aghabozorgi, S.; Shirkhorshidi, A.S.; Wah, T.Y. Time-series clustering—A decade review. *Inf. Syst.* **2015**, *53*, 16–38. [\[CrossRef\]](#)
- Ma, Z.; Yan, R.; Nord, N. A variation focused cluster analysis strategy to identify typical daily heating load profiles of higher education buildings. *Energy* **2017**, *134*, 90–102. [\[CrossRef\]](#)
- Li, K.; Yang, R.J.; Robinson, D.; Ma, J.; Ma, Z. An agglomerative hierarchical clustering-based strategy using Shared Nearest Neighbours and multiple dissimilarity measures to identify typical daily electricity usage profiles of university library buildings. *Energy* **2019**, *147*, 735–748. [\[CrossRef\]](#)
- Wu, X.; Kumar, V.; Quinlan, J.R.; Ghosh, J.; Yang, Q.; Motoda, H.; McLachlan, G.J.; Ng, A.; Liu, B.; Yu, P.S.; et al. Top 10 algorithms in data mining. *Knowl. Inf. Syst.* **2008**, *14*, 1–37. [\[CrossRef\]](#)
- Heidarinejad, M.; Dahlhausen, M.; McMahon, S.; Pyke, C.; Srebric, J. Cluster analysis of simulated energy use for LEED certified U.S. office buildings. *Energy Build.* **2014**, *85*, 86–97. [\[CrossRef\]](#)
- Deb, C.; Lee, S.E. Determining key variables influencing energy consumption in office buildings through cluster analysis of pre- and post-retrofit building data. *Energy Build.* **2018**, *159*, 228–245. [\[CrossRef\]](#)
- Dharssini, A.V.; Raja, S.C.; Karthick, T.; Venkatesh, P. Energy Pattern Classification and Prediction in an Educational Institution using Deep Learning Framework. *Electr. Power Compon. Syst.* **2022**, *50*, 615–635. [\[CrossRef\]](#)
- Liu, X.; Ding, Y.; Tang, H.; Xiao, F. A data mining-based framework for the identification of daily electricity usage patterns and anomaly detection in building electricity consumption data. *Energy Build.* **2021**, *231*, 110601. [\[CrossRef\]](#)
- Koupaei, D.M.; Cetin, K.; Passe, U.; Kimber, A.; Poleacovschi, C. Identifying rural high energy intensity residential buildings using metered data. *Energy Build.* **2023**, *298*, 113604.
- Paparrizos, J.; Gravano, L. k-Shape: Efficient and Accurate Clustering of Time Series. In *ACM SIGMOD Record*; Association for Computing Machinery: New York, NY, USA, 2015; Volume 45, pp. 69–76.
- Li, J.; Ma, R.; Deng, M.; Cao, X.; Wang, X.; Wang, X. A comparative study of clustering algorithms for intermittent heating demand considering time series. *Appl. Energy* **2024**, *353*, 122046. [\[CrossRef\]](#)
- Park, J.Y.; Yang, X.; Miller, C.; Arjunan, P.; Nagy, Z. Apples or oranges? Identification of fundamental load shape profiles for benchmarking buildings using a large and diverse dataset. *Appl. Energy* **2019**, *236*, 1280–1295. [\[CrossRef\]](#)
- Wen, L.; Zhou, K.; Yang, S. A shape-based clustering method for pattern recognition of residential electricity consumption. *J. Clean. Prod.* **2019**, *212*, 475–488. [\[CrossRef\]](#)
- Carmo, C.M.R.D.; Christensen, T.H. Cluster analysis of residential heat load profiles and the role of technical and household characteristics. *Energy Build.* **2016**, *125*, 171–180. [\[CrossRef\]](#)

26. Verleysen, M.; François, D. The Curse of Dimensionality in Data Mining and Time Series Prediction. In *Computational Intelligence and Bioinspired Systems, Proceedings of the IWANN 2005, Barcelona, Spain, 8–10 June 2005*; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2005; pp. 758–770.
27. Luo, X.; Hong, T.; Chen, Y.; Piette, M.A. Electric load shape benchmarking for small- and medium-sized commercial buildings. *Appl. Energy* **2017**, *204*, 715–725. [\[CrossRef\]](#)
28. Morris, B. The components of the Wired Spanning Forest are recurrent. *Probab. Theory Relat. Fields* **2003**, *125*, 259–265. [\[CrossRef\]](#)
29. Haben, S.; Singleton, C.; Grindrod, P. Analysis and Clustering of Residential Customers Energy Behavioral Demand Using Smart Meter Data. *IEEE Trans. Smart Grid* **2016**, *7*, 136–144. [\[CrossRef\]](#)
30. Yilmaz, S.; Chambers, J.; Patel, M. Comparison of clustering approaches for domestic electricity load profile characterisation—Implications for demand side management. *Energy* **2019**, *180*, 665–677. [\[CrossRef\]](#)
31. Hong, Y.; Yoon, S.; Choi, S. Operational signature-based symbolic hierarchical clustering for building energy, operation, and efficiency towards carbon neutrality. *Energy* **2023**, *265*, 126276. [\[CrossRef\]](#)
32. Kim, J.; Song, K.; Lee, G.; Lee, S. Time-series data clustering with load-shape preservation for identifying residential energy consumption behaviors. *Energy Build.* **2024**, *311*, 114130. [\[CrossRef\]](#)
33. Chen, S.; Lv, Y.; Wang, Z.; Ma, Y.; Huang, Y.; Wang, Y.; Cai, Y.; Rao, Z. Typical daily occupancy profiles of express hotels and its stochasticity effect on building heating and cooling loads. *J. Build. Eng.* **2023**, *73*, 106775. [\[CrossRef\]](#)
34. Ashouri, M.; Haghighat, F.; Fung, B.C.; Yoshino, H. Development of a ranking procedure for energy performance evaluation of buildings based on occupant behavior. *Energy Build.* **2019**, *183*, 659–671. [\[CrossRef\]](#)
35. Sun, C.; Zhang, R.; Sharples, S.; Han, Y.; Zhang, H. Thermal comfort, occupant control behaviour and performance gap—A study of office buildings in north-east China using data mining. *Build. Environ.* **2019**, *149*, 305–321. [\[CrossRef\]](#)
36. Wang, Y.; Shao, L. Understanding occupancy pattern and improving building energy efficiency through Wi-Fi based indoor positioning. *Build. Environ.* **2017**, *114*, 106–117. [\[CrossRef\]](#)
37. Wang, F.; Li, K.; Duić, N.; Mi, Z.; Hodge, B.M.S.; Shafie-khah, M.; Catalão, J.P. Association rule mining based quantitative analysis approach of household characteristics impacts on residential electricity consumption patterns. *Energy Convers. Manag.* **2018**, *171*, 839–854. [\[CrossRef\]](#)
38. Yu, Z.J.; Haghighat, F.; Fung, B.C.; Zhou, L. A novel methodology for knowledge discovery through mining associations between building operational data. *Energy Build.* **2012**, *47*, 430–440. [\[CrossRef\]](#)
39. Fan, C.; Xiao, F.; Madsen, H.; Wang, D. Temporal knowledge discovery in big BAS data for building energy management. *Energy Build.* **2015**, *109*, 75–89. [\[CrossRef\]](#)
40. Fan, C.; Xiao, F.; Song, M.; Wang, J. A graph mining-based methodology for discovering and visualizing high-level knowledge for building energy management. *Appl. Energy* **2019**, *251*, 113395. [\[CrossRef\]](#)
41. Zhang, C.; Zhao, Y.; Lu, J.; Li, T.; Zhang, X. Analytic hierarchy process-based fuzzy post mining method for operation anomaly detection of building energy systems. *Energy Build.* **2021**, *252*, 111426. [\[CrossRef\]](#)
42. Xu, Y.; Yan, C.; Shi, J.; Lu, Z.; Niu, X.; Jiang, Y.; Zhu, F. An anomaly detection and dynamic energy performance evaluation method for HVAC systems based on data mining. *Sustain. Energy Technol. Assess.* **2021**, *44*, 101092. [\[CrossRef\]](#)
43. Zhou, Y.; Yeoh, J.K.; Solihin, W. Studying the impact of building morphology on occupants’ movement using a rule mining approach. *Build. Environ.* **2024**, *249*, 111116. [\[CrossRef\]](#)
44. Sha, X.; Ma, Z.; Sethuvenkatraman, S.; Li, W. A novel rule mining method for knowledge discovery of relationships among indoor air quality, HVAC operation and occupants’ activities. *Build. Environ.* **2024**, *260*, 111670. [\[CrossRef\]](#)
45. Zhao, Y.; Zhang, C.; Cao, L. *Post-Mining of Association Rules: Techniques for Effective Knowledge Extraction*; IGI Global: Hershey, PA, USA, 2009.
46. Zhan, S.; Liu, Z.; Chong, A.; Yan, D. Building categorization revisited: A clustering-based approach to using smart meter data for building energy benchmarking. *Appl. Energy* **2020**, *269*, 114920. [\[CrossRef\]](#)
47. Zhang, C.; Zhao, Y.; Li, T.; Zhang, X. A post mining method for extracting value from massive amounts of post mining building operational data. *Energy Build.* **2020**, *223*, 110096. [\[CrossRef\]](#)
48. Hsu, D. Comparison of integrated clustering methods for accurate and stable prediction of building energy consumption data. *Appl. Energy* **2015**, *160*, 153–163. [\[CrossRef\]](#)
49. Rathod, R.R.; Garg, R.D. Regional electricity consumption analysis for consumers using data mining techniques and consumer meter reading data. *Electr. Power Energy Syst.* **2016**, *78*, 368–374. [\[CrossRef\]](#)
50. Fan, C.; Sun, Y.; Shan, K.; Xiao, F.; Wang, J. Discovering gradual patterns in building operations for improving building energy efficiency. *Appl. Energy* **2018**, *224*, 116–123. [\[CrossRef\]](#)
51. Gianniou, P.; Liu, X.; Heller, A.; Nielsen, P.S.; Rode, C. Clustering-based analysis for residential district heating data. *Energy Convers. Manag.* **2018**, *165*, 840–850. [\[CrossRef\]](#)
52. Dab, K.; Henao, N.; Nagarsheth, S.; Dubé, Y.; Sansregret, S.; Agbossou, K. Consensus-based time-series clustering approach to short-term load forecasting for residential electricity demand. *Energy Build.* **2023**, *299*, 113550. [\[CrossRef\]](#)
53. Choi, S.; Lim, H.; Lim, J.; Sungmin, Y. Retrofit building energy performance evaluation using an energy signature-based symbolic hierarchical clustering method. *Build. Environ.* **2024**, *251*, 111206. [\[CrossRef\]](#)

- 
54. Canaydin, A.; Fu, C.; Balint, A.; Khalil, M.; Miller, C.; Kazmi, H. Interpretable domain-informed and domain-agnostic features for supervised and unsupervised learning on building energy demand data. *Appl. Energy* **2024**, *360*, 122741. [[CrossRef](#)]
  55. Liu, Y.; Chong, W.T.; Yau, Y.H.; Wu, J.; Chang, Y.; Cui, T.; Chang, L.; Pan, S. A hybrid learning approach to model the diversity of window-opening behavior. *Build. Environ.* **2024**, *257*, 111525. [[CrossRef](#)]

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