energies

MDPI

Article

Electricity Pattern Analysis by Clustering Domestic Load Profiles Using Discrete Wavelet Transform

Senfeng Cen, Jae Hung Yoo and Chang Gyoon Lim *

Department of Computer Engineering, Chonnam National University, Yeosu 59626, Korea; jasoncsf.7@gmail.com (S.C.); jhy@jnu.ac.kr (J.H.Y.)
* Correspondence: cglim@jnu.ac.kr; Tel.: +82-61-659-7254

Abstract: Energy demand has grown explosively in recent years, leading to increased attention of energy efficiency (EE) research. Demand response (DR) programs were designed to help power management entities meet energy balance and change end-user electricity usage. Advanced real-time meters (RTM) collect a large amount of fine-granular electric consumption data, which contain valuable information. Understanding the energy consumption patterns for different end users can support demand side management (DSM). This study proposed clustering algorithms to segment consumers and obtain the representative load patterns based on diurnal load profiles. First, the proposed method uses discrete wavelet transform (DWT) to extract features from daily electricity consumption data. Second, the extracted features are reconstructed using a statistical method, combined with Pearson's correlation coefficient and principal component analysis (PCA) for dimensionality reduction. Lastly, three clustering algorithms are employed to segment daily load curves and select the most appropriate algorithm. We experimented our method on the Manhattan dataset and the results indicated that clustering algorithms, combined with discrete wavelet transform, improve the clustering performance. Additionally, we discussed the clustering result and load pattern analysis of the dataset with respect to the electricity pattern.

Keywords: demand response; discrete wavelet transform; Pearson's correlation coefficient; principal component analysis; clustering

Citation: Cen, S.; Yoo, J.H.; Lim, C.G. Electricity Pattern Analysis by Clustering Domestic Load Profiles Using Discrete Wavelet Transform. *Energies* **2022**, *15*, 1350. https://doi.org/10.3390/en15041350

Academic Editors: Sergio Nesmachnow and Islam Safak Bayram

Received: 10 December 2021
Accepted: 11 February 2022
Published: 13 February 2022

1. Introduction

Smart grid technologies and applications capable of adaptive, resilient, and sustainable self-healing, with foresight for prediction under different uncertainties, improve the reliability of the power system [1]. Furthermore, the smart grid allows bidirectional communication that supports the demand response (DR) programs [2]. Demand response technologies are widely applied and are constantly improving. The most common DR programs can be categorized into the following two classes: price-based programs and incentive-based programs. Price-based programs contain time of use (ToU), real time pricing (RTP) and critical peak pricing (CPP), which aim to motivate the end-user to change their consumption behavior [3]. On the other hand, incentive-based programs reach a consensus with consumers to reduce electricity consumption. Examples of these schemes are direct-load control (DLC), interruptible/curtailable service (I/C), demand bidding/buy (DB), etc. [4]. Considering various end-user consumption behaviors, it required the utility companies to design reasonable strategies. Therefore, it is necessary to analyze end-users' consumption data to acquire the load patterns.

Advanced metering infrastructure (AMI) and smart meters have been adopted to automatically collect energy consumption data at a fine granular interval, which is usually in intervals of 1 h, 30 min, or even 30 s [5]. Most countries have vigorously deployed smart meters because of the potential value of consumption data [6]. The massive amount of data sampled by smart meters could be used for research, typically load forecasting, customer segmentation, pricing/incentive mechanism, scheduling and control [7].

However, the extracted load consumption data lack labels, hence, the need of clustering techniques to segment the electricity consumption data. In addition, with the high time resolution advanced smart meter implemented in the household, the massive data will increase the complexity of the clustering method, called the "curse of dimensionality" [8]. This is a problem for implementing clustering algorithms because most clustering algorithms become intractable to process high-dimensional data input. To deal with the issue of the curse of dimensionality, the load consumption data needs preprocessing i.e., dimensionality reduction.

This study proposed clustering for segment residential customer daily power data, using discrete wavelet transform to extract features and reduce dimension by statistical methods and principal component analysis (PCA). The dataset, named Multifamily Residential Electricity Dataset (MFRED), contains 10-s resolution daily power data for 26 apartment groups, collected over 365 days in Manhattan, New York, 2019 [9]. First, data cleansing and multi-level one-dimensional (1D) discrete wavelet transform were applied on 8640-value daily load curves. Second, we reduced extracted feature dimensions. Finally, clustering algorithms were implemented, and the evaluation of the methods was carried out. Our main contributions of this work include the following: (1) a proposed method to vastly reduce the daily load profile dimensionality, to accelerate the clustering, and (2) the three cluster validity indices (CVI) imply that our proposed method to extract features outperforms the clustering original data, especially on hierarchical clustering.

The paper is structured as follows: Section 2 briefly discusses the related works. Section 3 describes the MFRED data. Section 4 explains the methodology in the study. Analysis and results are presented in Section 5, with conclusions in Section 6.

2. Related Works

Clustering is unsupervised learning, which could group similar data with no label attached to them [10]. Clustering algorithms can be classified into partitioning algorithms, hierarchical algorithms, density-based algorithms, and grid-based algorithms [11]. The authors of [12] implemented an improved K-means clustering method on load curves and verified that it performed better than the original K-means algorithm. The authors in [13] used modified fuzzy c-means (FCM) to extract representative load profiles of the customers. Ordering points to identify the clustering structure (OPTICS) is one of the density-based clustering models used to analyze consumer bid-offers in [14]. Gaussian mixture model (GMM) clustering is widely used to segment households' load profiles for demand response [15].

Additionally, most clustering algorithms cannot properly process high dimensionality data [16]. Most of the aforementioned works extracted consumption load patterns in terms of hourly, 30-min, 15-min load data. However, the advanced high-frequent smart meter could extract load data in intervals of 1-min, 30-s, and even 1-s, leading to large-scale consumption data that increases computational complexity. Most clustering algorithms evaluating the belonged cluster are calculated by distance. High dimensionality data would consume more computational complexity in each iteration, resulting in more time consumption. Hence, there are numerous studies about dimensionality reduction on load curve clustering, using feature extraction, feature construction and feature selection. In [17], the authors developed electricity price schemes based on demand patterns, using k-means combined with PCA. In [18], the authors proposed singular value decomposition to extract features before k-means clustering and evaluate the error sum of squares (SSE) index to compare with direct clustering. In [19], they used a fused load curve k-means algorithm, based on "Haar" discrete wavelet transform for reduce dimension, to obtain the load patterns of consumers from China and the United States and evaluate clustering performance by four CVI [20]. Xiao et al. [21] proposed a fusion clustering algorithm to obtain the consumption characteristics, using load curve clustering, based on discrete wavelet transform (CC-DWT).

In this study, we implemented clustering to segment 10-s interval daily electricity consumption data, using multi-level discrete wavelet transform, Pearson correlation coefficient, and PCA techniques to preprocess the daily load profiles. The clustering evaluation result shows our proposed method outperformed the conventional methods, without reducing dimension.

3. Data

In this study, we used the Multifamily Residential Electricity Dataset (MFRED) [9], which consisted of 390 apartments, from 1 January to 31 December 2019. This dataset was collected by real-time metering and contained 246 million data from residential buildings in Manhattan, New York, USA. The resolution of data was one sample per 10-s, providing 8640 data points in each daily profile. During the one-year period, some advanced meters were offline due to various reasons (e.g., smart meters offline). Therefore, some electricity data were not recorded in MFRED.

In the MFRED, the percentages of building stock prior to 1940, between 1940–1980, post-1980 were 79%, 7%, and 14%, respectively. The ratios of the entire Manhattan building stock prior to 1940, between 1940–1980, post-1980 were 86%, 6%, 8%, respectively, which means the residential structure in our research is very similar to that of the whole of Manhattan. In addition, considering the privacy leakage, the 390 apartments' data were reconstructed into 26 groups, called apartment groups (AG), which means each AG is made up of 15 apartments that are more representative. Hence, the dataset recorded the average real power (kW), reactive power (kVAR) and consumption (kWh), over 15 apartments, from 26 apartment groups, every 10 s for 365 days. Here, we used one channel real power data for our research. Figure 1 shows the distribution of daily energy consumption, and the black dashed line represents the mean electricity consumption (8.21 kWh).

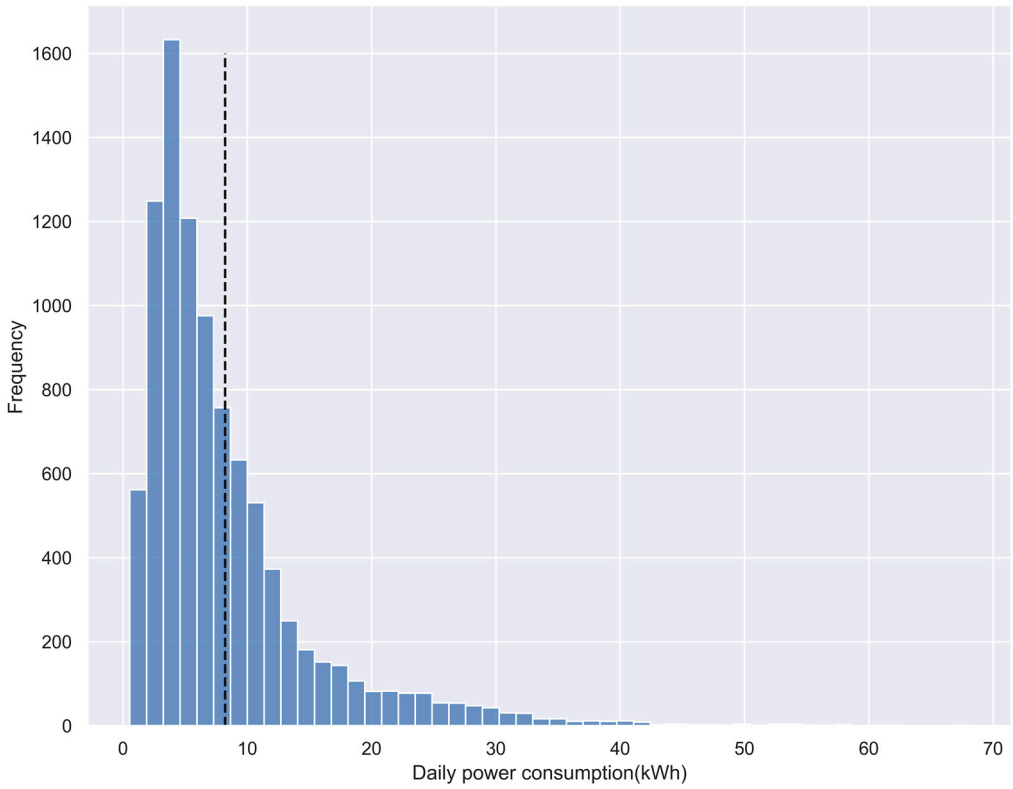

Figure 1. Daily energy consumption distribution.

4. Methodology

Our proposed method consists of the following four major stages: data cleansing, feature extraction, dimensionality reduction and clustering. Daily real power data are obtained from MFRED, and the data are cleansed for the missing value. Multi-level discrete wavelet transform is then applied to extract the features. In the dimensionality reduction stage, we implement the following two methods to decrease the dimension: statistical method combined with Pearson correlation and PCA. Finally, clustering algorithms were applied to segment daily load curves by using selected features. The proposed method is as shown in Figure 2.

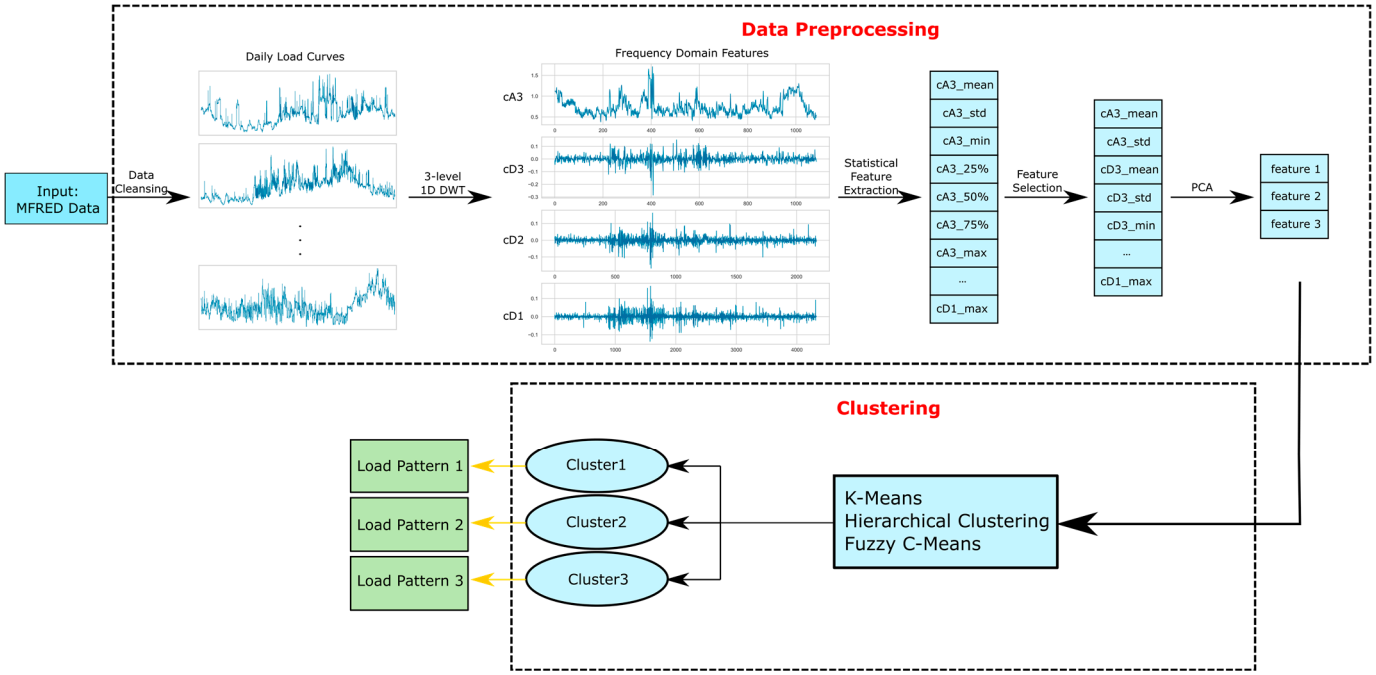

Figure 2. The proposed system diagram for electricity pattern analysis by clustering domestic load profiles.

4.1. Data Cleansing

Real and reactive power data were recorded in MFRED, where the real power data is reserved for the purpose of clustering. The primary issue with real power data is the missing values and anomalous values. Missing values are filled by averaging the previous and post 10-s values. However, tens of thousands of continuous data were missed because of the long-time breakdown of all meters on 09 July 2019, from 14:30 to 21:30 UTC. Therefore, this day is not taken into consideration due to the large amount of missing data. Anomalous values may be caused by the real-time meters (RTM) data collection accuracy, detected by the following five-number summary: the minimum, the maximum, the sample median, and the first and third quartiles. The single outlier was replaced by the average, the maximum and itself. After data cleansing, the reconstructed subset consisted of 8640 ten-second interval real power data (kW) in 364 days and 26 AGs. Thus resulting input data matrix dimension is 9464 $\times$ 8640. Figure 3 illustrates the 8640-value diurnal load curves from different AGs.

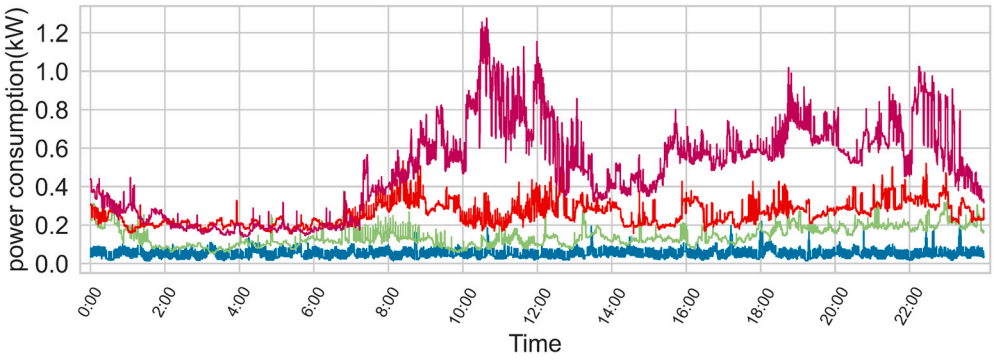

Figure 3. Daily load curves from different AGs. Each daily load curve consisted of 8640 values from meters every 10 s from 00:00:00 to 23:59:50.

4.2. Discrete Wavelet Transform

Wavelet transform contains continuous wavelet transform (CWT) and discrete wavelet transform (DWT). Discrete wavelet transform is widely used in waveform processing, including feature extraction in electroencephalography (EEG) [22], electromyography (EMG) [23], time-series load curves [24,25], etc. DWT decomposes the signal into various sets by passing through the low-pass filter and high-pass filter. The DWT and DWT coefficients are given by Equations (1) and (2), respectively, as follows:

$$\psi_{j,k}(t) = 2^{-\frac{j}{2}}\psi\left(2^{-j}t - k\right) \tag{1}$$

$$W_{j,k} = W\left(2^j, k2^j\right) = 2^{-j/2}\int_{-\infty}^{\infty}\overline{\psi\left(2^{-j}t - k\right)}dt \tag{2}$$

where k is a signal index and j is the scale index.

The detailed coefficients are obtained from a high-pass filter, while approximation coefficients are extracted from a low-pass filter, which could continue to decompose into a high-pass filter and low-pass filter. Figure 4 shows the decomposition of the 3-level 1-D discrete wavelet transform that we used in our research.

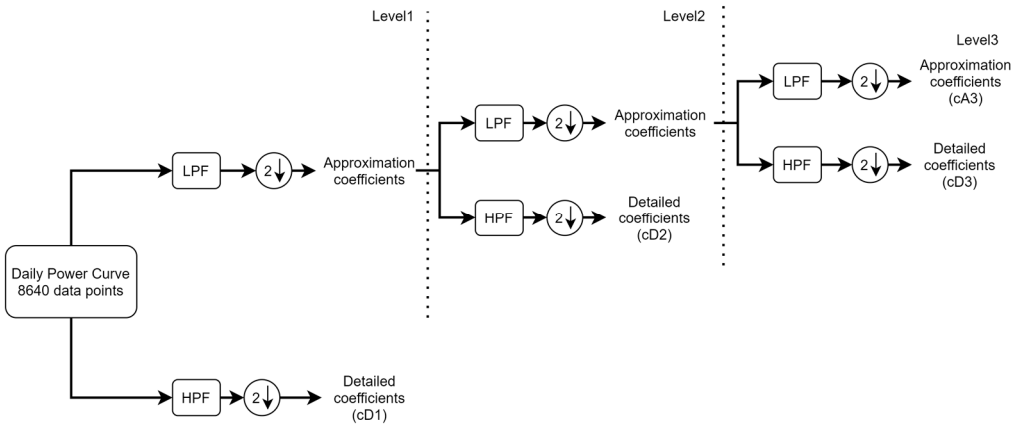

Figure 4. Diagram of the multi-level 1-D discrete wavelet transform.

To extract features from daily load curves, we implemented the three-level 1-D Daubechies 4 (db4) discrete wavelet transform. Three-level means it will repeat one-level 1-D discrete wavelet transform three times based on the previous approximation coefficients. The Daubechies wavelet is preferred for feature extraction compared with Haar wavelet which is the special case of Daubechies noted as db1. Haar wavelet is the simplest and first wavelet transform which decomposes the discrete data using the two-length filter.

Eight of filter length in db4 wavelet contains more details but it involves slightly higher computational processes [19]. Thus, we employed db4 to compute the detailed coefficients and approximation coefficients. Three detailed coefficient sets and one approximation coefficient set are denoted as cD1, cD2, cD3, cA3, respectively. Figure 5 shows the four components of the daily load curve while using a three-level 1-D db4 discrete wavelet transform. The cA3 coefficients curve reflects a similar variation with the original load curve, while the value of cD3, cD2, cD1 components is very close to 0, which contains detailed information of daily load curve. For each daily power curve, the number of detailed coefficients (cD1, cD2, cD3) and approximation coefficients (cA3) were 4323, 2165, 1086 and 1086, respectively.

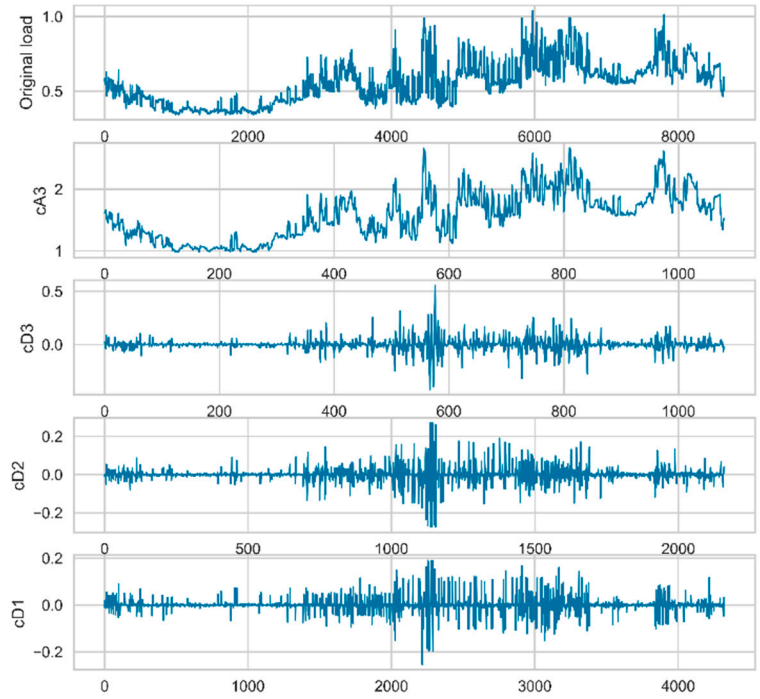

Figure 5. The curves of the components using the db4 wavelet.

4.3. Dimensionality Reduction

This phase aims to reduce the dimensions from extracted features (cA3, cD3, cD2, cD1). First, we used the statistical method to get the statistical variables (mean, std, min, 25%, 50%, 75%, max) from each daily coefficient (cA3_mean, cA3_std, cA3_min, cA3_25%, cA3_50%, cA3_75%, cA3_max, cD3_mean, etc.). There were 28 features extracted from the approximation and detailed coefficients. Second, we calculated Pearson's correlation coefficient, which measures the correlation of each two features. The correlation coefficient values are between -1 and 1, the value close to 1 represents a high positive correlation while the value close to -1 represents a high negative correlation [26]. High correlation features can be replaced by other features with similar characteristics. The correlation coefficient value is calculated from Equation (3), as follows:

$$r_{XY} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}} \quad (3)$$

where n is the number of samples, X_i, Y_i is the value of data, $\overline{X}$ is the mean value of X, and $\overline{Y}$ is the mean value of Y.

The correlation heatmap that represents the coefficient matrix is shown in Figure 6. According to the correlation heatmap, coefficients close to 1 or -1 imply redundant features. For the purpose of reducing the dimension, we removed one of the features in which

the absolute values of correlation coefficients are bigger than 0.95. Figure 7 shows the correlation heatmap after eliminating the high correlation features.

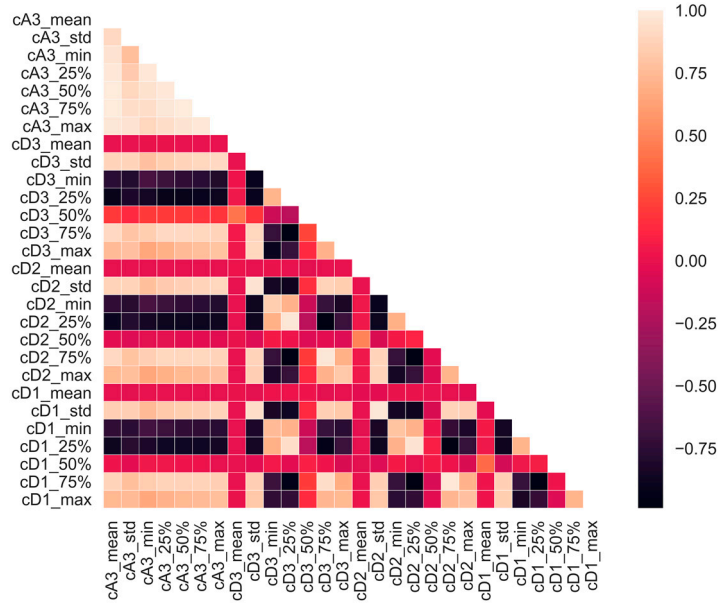

Figure 6. Correlation heatmap representing the coefficient matrix of 28 features.

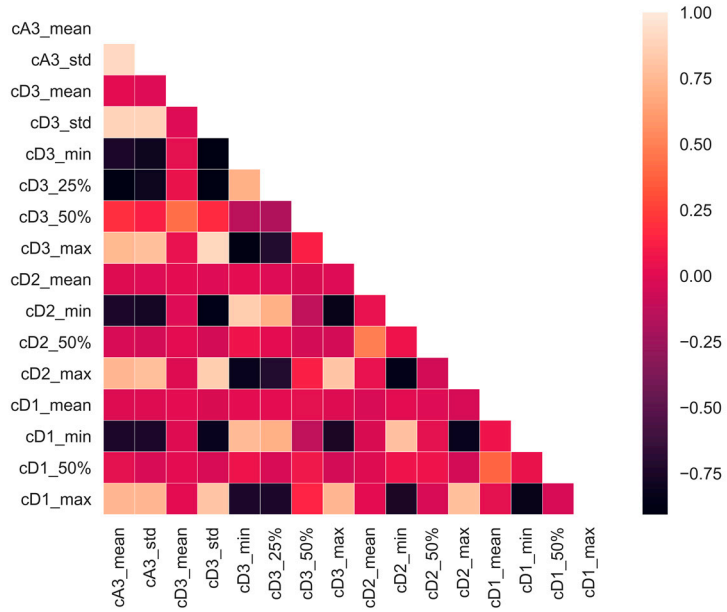

Figure 7. The selected features correlation heatmap after eliminating 12 highly correlated features.

PCA is one of the most appealing techniques that is widely used for dimensionality reduction of large data sets [27]. Given the original high dimensional data, PCA can map the data into k dimensions ($k <$ original dimension) with principal components that are not related to each other and still preserve the original information. The data are normalized by z in the PCA process to obtain the feature vector composed of principal components by finding the covariance, eigenvector, and eigenvalue. In our study, we applied the PCA method to reduce the dimensionality to 3 components and still preserve 99 percent variability. Thus input dimension is reduced to 28 from 8640 using DWT combined with the statistical method and to 16 after correlation analysis. Finally, we have 3 component features applying PCA transform.

4.4. Clustering Method

Clustering load curves into groups is essential to identify load patterns [28]. For comparison purposes, we implemented the following three different clustering methods: k-means, hierarchical, and fuzzy c-means clustering.

4.4.1. K-Means Clustering

K-means is the most popular hard clustering algorithm goal to partition n data into k clusters, with the grouped data close to its centroid [29]. The K-means method is implemented as follows: First, determine the cluster number K then initial K centroids. Second, allocate each sample to the nearest centroids according to the distance. Third, determine the new K centroids that were generated by calculating the mean of the cluster points. Then, repeat the second and third steps until the centroids are completely unchanged.

Determining the number of clusters is one of the major challenges in clustering. The elbow method aims at finding the appropriate number of clusters by calculating the score for a range of values of K [30]. In our study, we determined this parameter by analyzing the following two metrics: distortion and Calinski–Harabasz score. Generally, distortion scoring computes the within cluster sum of squared (WCSS) to select cluster K [31]. Distortion score decreases with K increase. It is computed using Equation (4), as follows:

$$WCSS(K) = \sum_{h=1}^{K} \sum_{x_i \in c_h} \|x_i - \mu_h\|^2 \tag{4}$$

where K is the number of clusters, c_h is the cluster h, μ_h is the hth cluster center, $\|x_i - \mu_h\|^2$ is the Euclidean distance between data point x_i and its belonged centroid μ_h.

We applied Yellowbrick package to visualize the elbow method [32]. Figure 8 illustrates the WCSS value in different K. By applying the elbow method for $1 \leq K \leq 10$, the distortion score reduces rapidly with increase in K until $K = 3$ and then reduces gradually. We also employed Calinski–Harabasz analysis method in our study. It calculates the ratio of the sum of between-clusters dispersion and inter-cluster dispersion for all clusters, as follows:

$$CH(K) = \frac{\sum_{h=1}^{K} n_h \|c_h - c\|^2}{\sum_{h=1}^{K} \sum_{x^{(i)} \in c_h} \|x_i - c_h\|^2} \frac{N-K}{K-1} \tag{5}$$

where N is the total number of data points, K is the number of clusters, n_h and c_h are the number of points and centroids of the hth cluster, respectively, c is the centroid of data points. The higher value of $\sum_{h=1}^{K} n_h \|c_h - c\|^2$ means different cluster centroids are well separated, while the lower value of $\sum_{h=1}^{K} \sum_{x^{(i)} \in c_h} \|x_i - c_h\|^2$ indicates that the points of cluster are well centered. Therefore, the larger the value of the CH index, the more distinct the clusters.

Figure 9 shows the scores according to the change in the value of K, and it has a maximum value when $K = 3$. Even looking at the graph combined with the distortion and Calinski–Harabasz scores, it proves that it is the optimal solution for $k = 3$.

4.4.2. Hierarchical Clustering

Hierarchical clustering algorithms are formed by iteratively dividing the groups using bottom-up or top-down methods called agglomerative and divisive hierarchical clustering [33]. In this study, we employed agglomerative hierarchical clustering to segment load curves based on preprocessed features. The agglomerative builds up clusters starting with a single object as a single cluster and then using distance metric to merge the two most similar clusters [34]. Repeat until all of the objects are finally merged into a single cluster. We use "Ward" linkage to compute the distance between the new cluster and the rest of the

clusters, minimizing the variance of the merged clusters [35]. Ward linkage criterion can be expressed as follows:

$$\Delta(X_i, X_j) = \frac{n_i n_j}{n_i + n_j} \| c(X_i) - c(X_j) \|^2 \quad (6)$$

where $c(X_i)$ is the centroid of cluster i, n_i denotes the number of points in cluster i.

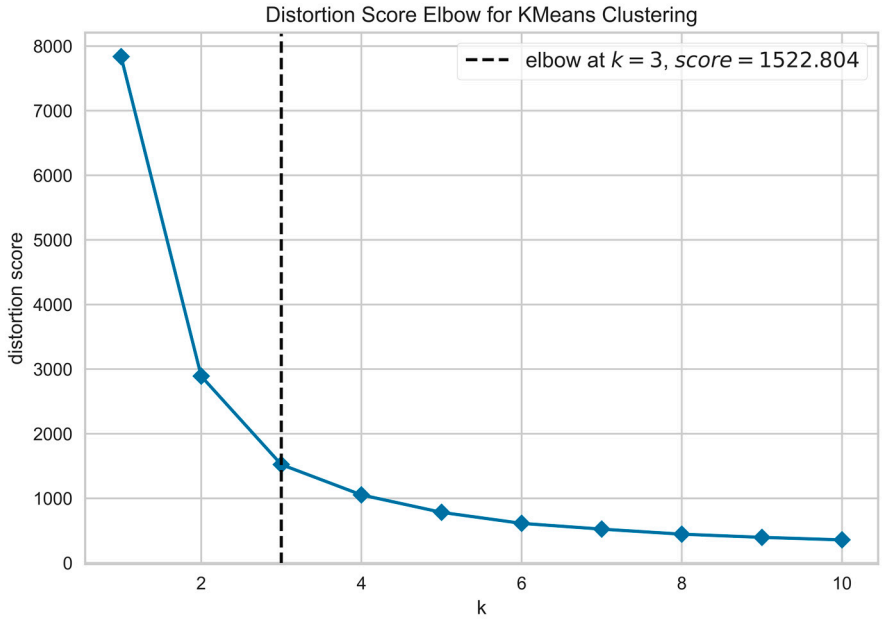

Figure 8. Elbow method estimated by distortion.

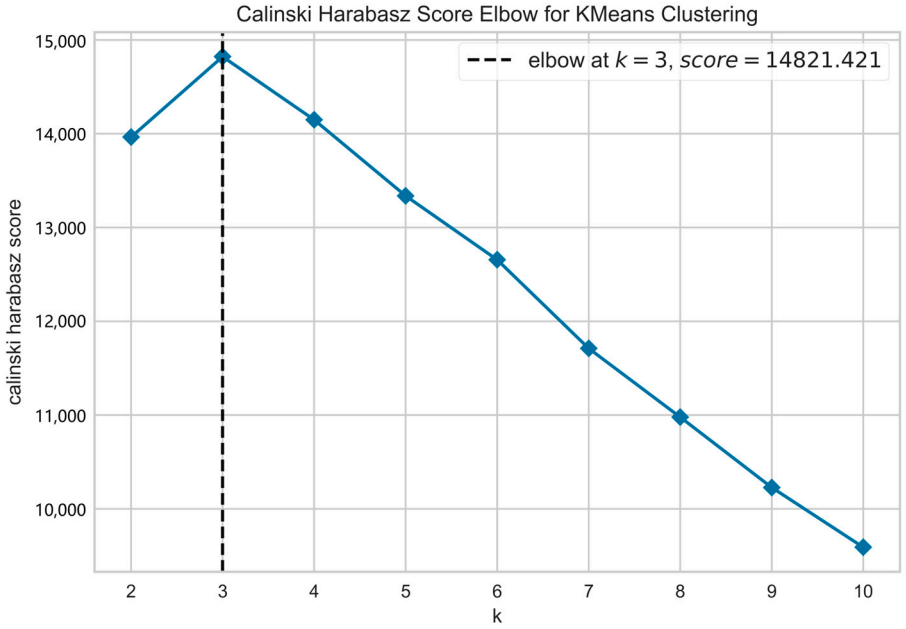

Figure 9. Elbow method estimated by Calinski–Harabasz.

Figure 10 depicts the Ward linkage truncated dendrogram which present a tree structure to vasualize the clusters and the number belonged each cluster. Ward's method dendrogram displays the clustering structure of the data. The numerical data in Figure 10 means the distance between different cluster centers which is calculated by Equation (6). The black dashed line represents the distance threshold which is 50. In addition, we combined the Calinski–Harabasz index with dendrogram to determine the optimal number of clusters (Table 1). According to the result, it can be confirmed that when K changes from 2 to 3, the Calinski–Harabasz index increases rapidly and then gradually increases thereafter. The Calinski–Harabasz index and dendrogram indicate that three is the optimal number for the value of K.

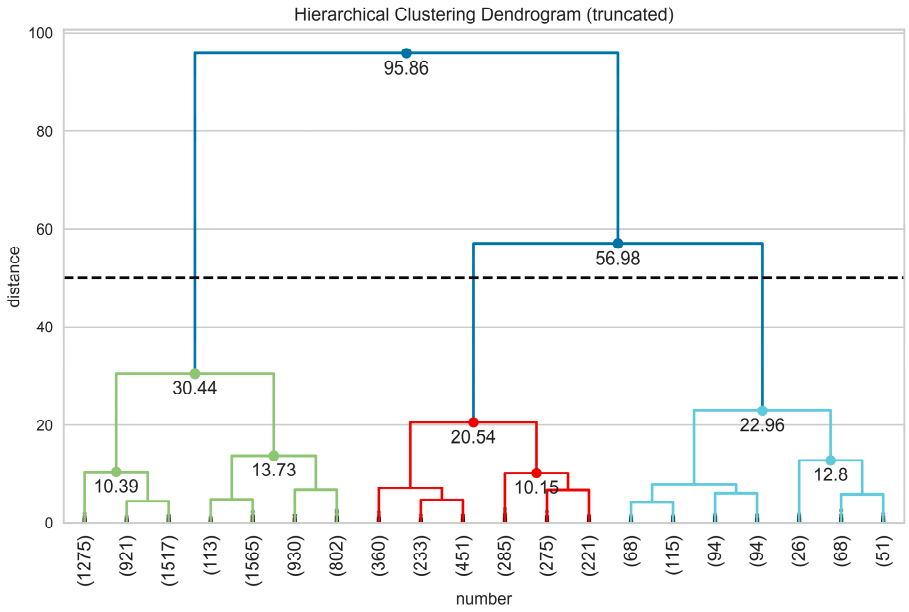

Figure 10. Agglomerative clustering dendrogram using Ward linkage.

Table 1. Calinski–Harabasz Index of agglomerative clustering.

Cluster(K)	Calinski–Harabasz Index
2	13,409.27
3	18,170.62
4	18,235.64
5	18,414.97
6	19,879.79
7	19,482.06
8	19,626.11
9	19,368.50

4.4.3. Fuzzy c-Means Clustering

The fuzzy c-means (FCM) algorithm is one of the soft clustering algorithms, also known as "soft K-means," where each data object can belong to multiple clusters. The fuzzy c-means algorithm has been widely used in many applications, such as consumer behavior and market segmentation [36]. FCM aims to minimize the objective function, as follows:

$$J_m = \sum_{i=1}^{N} \sum_{j=1}^{C} u_{ij}^{m} \|x_i - c_j\|^2 \quad (7)$$

where m is the fuzziness parameter in the range of [1,+∞), u_{ij} is the degree of membership of x_i in cluster j, c_j is the centroid of cluster j. The membership degree and cluster center will be updated iteratively until the objective function value is smaller than the error. The cluster center c_j and membership degree u_{ij} and can be obtained as follows:

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^m x_i}{\sum_{i=1}^{N} u_{ij}^m} \tag{8}$$

$$u_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|} \right)^{\frac{2}{m-1}}} \tag{9}$$

The algorithm comprises the following steps:

Step 1: Determine the number of clusters, fuzziness parameter m and the error ε.

Step 2: Initialize the membership matrix $U^{[0]}$ using $\sum_{j=1}^{c} \mu_j(x_i) = 1$.

Step 3: At k step, compute the centroid c_k with equation (8).

Step 4: Update the new membership matrix $U^{[k]}, U^{[k+1]}$ with Equation (9).

Step 5: If $\|U^{[k+1]} - U^{[k]}\| < \varepsilon$, stop, else, return to step 3.

The main advantage of FCM is its suitability for overlapped data, its scalability and simplicity, and accuracy. However, the time complexity of fuzzy c-means is more than *k-means*. In our study, we selected the fuzziness index and error ε by grid search. The optimal fuzziness index was determined as $m = 1.25$, and the error as $\varepsilon = 1 \times 10^{-5}$. Figure 11 shows the clustering result based on three principal components. The points from Cluster 1 and Cluster 2 are relatively compact, but Cluster 3 is more dispersed.

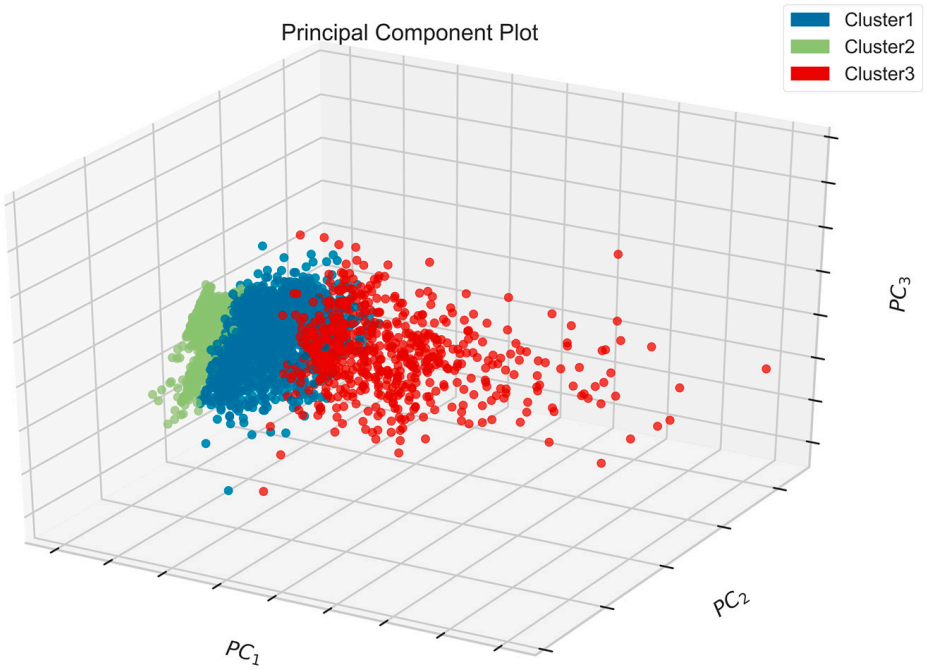

Figure 11. Clustering result with FCM applied when c is set to 3.

5. Experiment Results and Analysis

In the clustering phase, we employed three different clustering algorithms to segment different daily load curves. Considering the diversity of clustering performance evaluations, we selected three indices for validation, namely the silhouette coefficient, Calinski–Harabasz index, and Davies–Bouldin index (DBI) [37], which are internal clustering criteria. The Calinski–Harabasz index has been described in Section 4. The silhouette coefficient combines cohesion and separation. Cohesion indicates the similarity of points in the same cluster. On the contrary, separation indicates the object compared to other clusters. Specifically, the silhouette coefficient is calculated as follows:

$$SC = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{10}$$

where *a(i)* indicates that cohesion is the mean distance between a sample and all other points in the same cluster, and *b(i)* is the minimum value of the mean distance between an object and all other objects in the nearest cluster, then, the equations of *a(i)* and *b(i)* are as follows:

$$a(i) = \frac{1}{|C_i| - 1} \sum_{j \in C_i, i \neq j} d(i, j) \tag{11}$$

$$b(i) = \min_{k \neq i} \frac{1}{|C_k|} \sum_{j \in C_k} d(i, j) \tag{12}$$

The value of silhouette is in the range of [−1,1]. If the silhouette coefficient is close to 1, it means that the model is suitable; a negative value indicates incorrect clustering. Higher values of the silhouette coefficient imply that the model clustered well. Davies–Bouldin index measures the average similarity between clusters, where the similarity compares the distance between clusters with the size of the clusters themselves. For a given set of clusters $C = \{c_1, c_2, \ldots, c_k\}$, c_i is the most similar with c_j. Davies–Bouldin index is defined as follows:

$$DB = \frac{1}{k} \sum_{i=1}^{k} \max_{i \neq j} \frac{s_i + s_j}{d_{ij}} \tag{13}$$

where k is the cluster number, s_i is the average distance between all objects in cluster i and cluster i centroid, d_{ij} is the distance between *ith* and *jth* cluster centroids. The smaller value of the Davies–Bouldin index implies that the clusters are separated properly.

We compared our proposed method with the original clustering algorithm without reducing the dimension. Table 2 compares the three clustering results, presented by calculating cluster validity indexes. The name of clustering methods that include 'Original' denotes the daily load data without reducing dimensionality. *N* denotes that the daily load data were normalized by min–max normalization to rescale the data to fit in the range 0 to 1. Generally, normalizing the data before clustering could ignore the distance difference between different variables. Equation (14) presents the formula for min–max normalization, as follows:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{14}$$

According to the evaluation index, our wavelet-based preprocessing method slightly improves clustering performance compared to the original method. However, the performance of wavelet-based hierarchical clustering is better than hierarchical clustering without dimensionality reduction. Compared with the three wavelet-based clustering algorithms, the performance of k-means and FCM were similar, the silhouette coefficient and Davies–Bouldin index of FCM were better than k-means. For hierarchical clustering, the silhouette coefficient is the best, but the other two indices are worse than those of k-means and FCM. In addition, the proposed method significantly saves the computation time by dimensionality reduction.

Table 2. Clustering evaluation comparison results of the proposed methods.

Methods	SC	CH	DBI
Wavelet based K-means	0.5101	13,643.7	0.7741
Wavelet based HC	0.5351	12,795.9	0.7909
Wavelet based FCM	0.5105	13,642.4	0.7736
Original_N_K-means	0.5103	13,642.4	0.7758
Original_N_HC	0.4105	11,431.5	0.8367
Original_N_FCM	0.5058	13,609.7	0.7759

Note: SC is Silhouette Coefficient, CH is Calinski–Harabasz Index, and DBI is Davies–Bouldin index. The larger the SC and CH values, the better. Conversely, the smaller the DBI values, the better.

Based on our comparison, we adopt the wavelet-based fuzzy c-means method. In three clusters, the first, second, and third clusters represent 66.54%, 26.84%, and 6.62% of the daily load curves, respectively. Figure 12 shows the load patterns of the three clusters and daily load curves. Cluster 1 and 3 represent the lowest and highest power consumption, respectively. In each cluster figure, the bold red line represents the representative load pattern, while the other curves represent the daily power usage in the cluster. Cluster 1 contains 6297 daily load curves, with stable power consumption; the average power usage and average peak power were 0.187 kW and 0.438 kW, respectively. Cluster 2 contains 2540 daily load curves; the average power usage and average peak power were 0.517 kW and 1.056 kW, respectively. Cluster 3 is composed of 627 daily load curves, which is the highest power usage group and has the highest variability. For Cluster 3, the average power usage and average peak power were 1.212 kW and 2.209 kW, respectively.

Figure 13 illustrates the average daily power usage box and whisker plot of three clusters. Boxplot could present data distribution based on a five-number summary, including minimum, first quartile, median, third quartile and maximum. There are some outliers (data point in Figure 13) in cluster 2, while in cluster 3, many outliers fall beyond the maximum value. As the power usage increases from cluster 1 to 3, the variation in power also increases. The standard deviation of cluster 1 is 0.0829 kW, cluster 2 is 0.1329 kW, and cluster 3 is 0.3193 kW.

Figure 14 shows the average power load pattern in four seasons of three clusters. It appears that the three clusters have similar power usage characteristics in the four seasons, i.e., the average power usage valley and peak at the same time every season, around 4 am and 8 pm, respectively. Moreover, the household generally needs to use air conditioners to control the indoor temperature during the summer; therefore, electricity usage is higher. The winter consumption in the three clusters is less than that of the summer, insinuating that most apartments have installed a heating system that is not taken into account in the electricity data. Looking at all four seasons, electricity demand is stably required between 8 am and 2 pm in Cluster 1 (a) and Cluster 2 (b). The section that consumes the most power is Cluster 3, and it can be seen that the power demand increases over time during the same period.

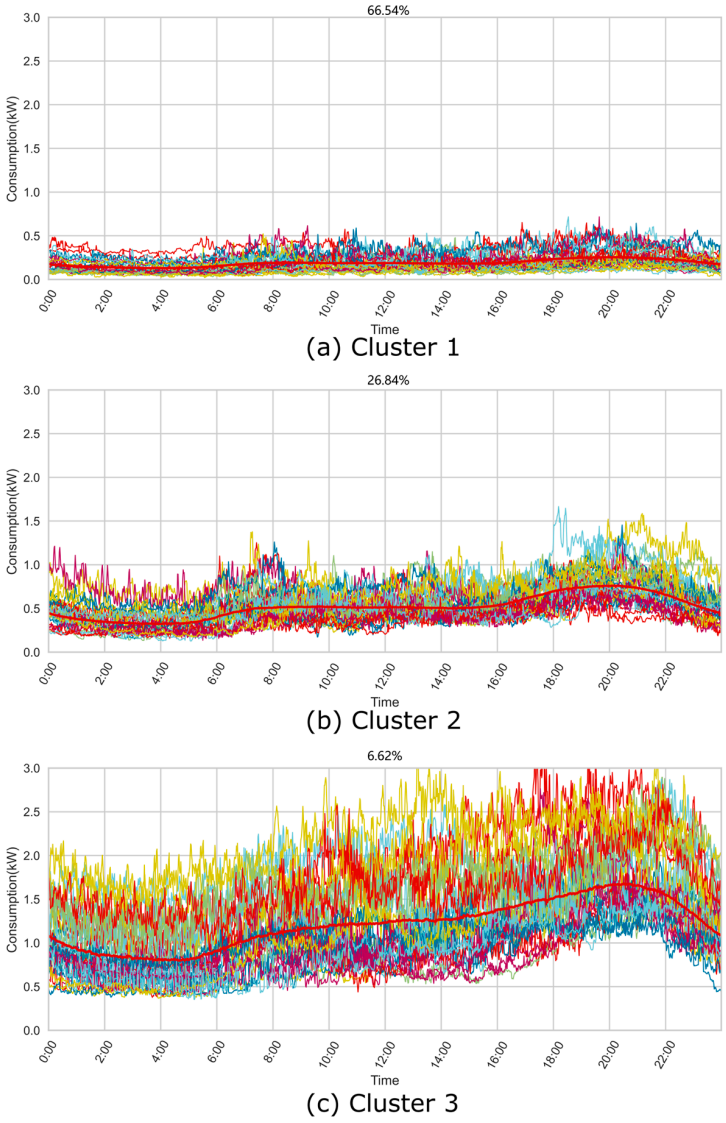

Figure 12. Daily load curves and load patterns of each cluster, (**a**) low load consumption group, (**b**) middle load consumption group, and (**c**) high load consumption and instability group.

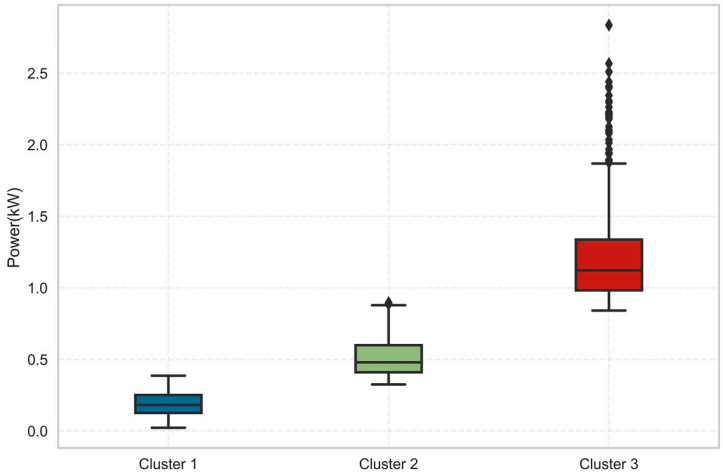

Figure 13. Box and whisker plot of average daily power (kW) usage in three clusters.

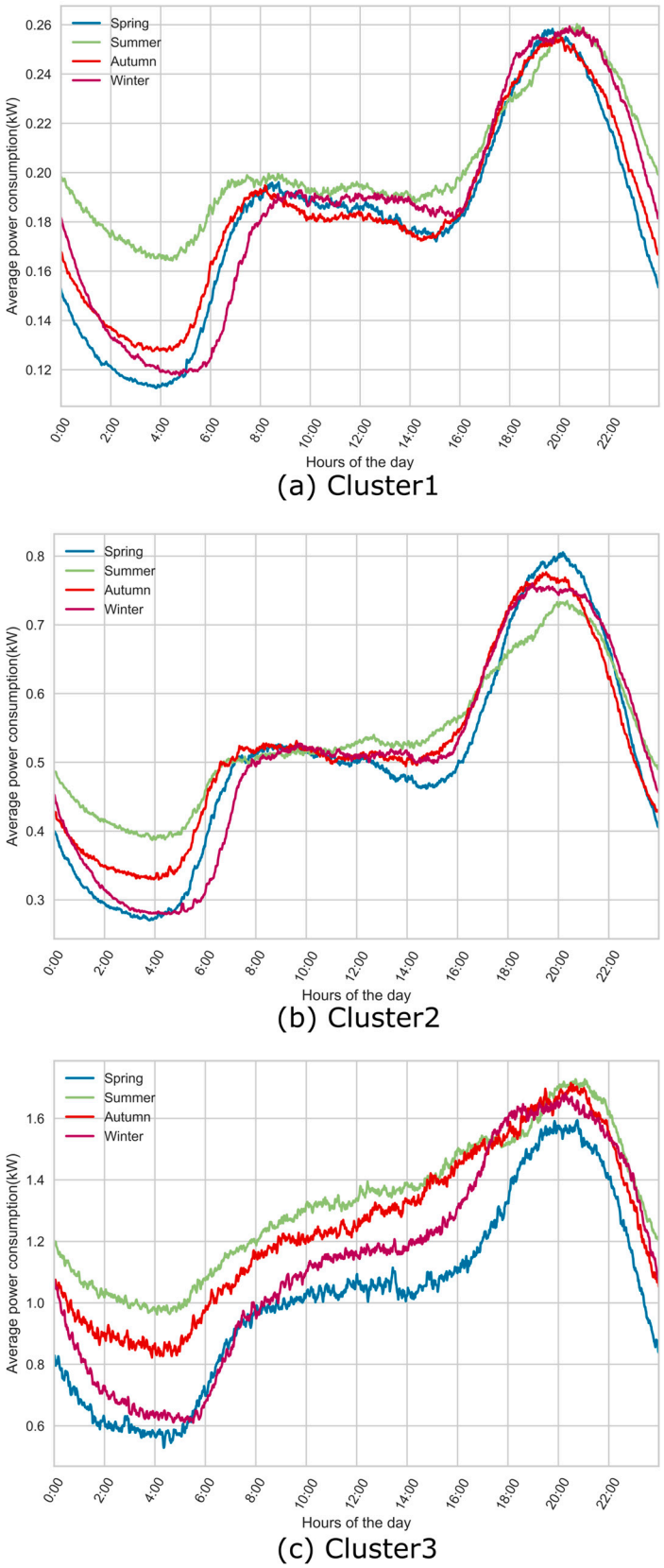

Figure 14. Seasonal average load patterns in three clusters, (**a**) low load consumption group, (**b**) middle load consumption group, and (**c**) high load consumption and instability group in four seasons.

6. Conclusions

High-frequency smart meters have been broadly deployed for collecting electricity data. Our proposed method implements discrete wavelet transform to convert time-domain data to frequency domain. We extracted detailed and approximate signals using a statistical approach, then used Pearson's correlation coefficients to filter the high correlation features. To further reduce dimensionality, we applied PCA to preserve three features. The rest of the three features were used to achieve the clustering algorithm. Our study aimed at obtaining the representative load patterns from high time resolution daily load curves in Manhattan. Our method reduces the large dimensions to increase efficiency in clustering. In addition, it improves the clustering result slightly by estimating the silhouette coefficient, Calinski–Harabasz index, and Davies–Bouldin index, then comparing the clustering without discrete wavelet transform. From representative load patterns, the utility policymaker could design a reasonable demand response scheme to maintain the power system stability and help the utility maximize the profit and even reduce consumers' electricity fees. Based on Figure 14, policymakers could design three different advanced time of use tariffs, according to electricity consumption volume and representitive load curves from the three clusters. To each cluster, the electricity demand increases apparently from 4 pm to 8 pm, which could influence the power system stability. It means the appropriate DR scheme is significant during this period, such as load shifting/shedding. For future work, we suggest exploring the sub-cluster from the previous clusters to get more detailed load patterns based on our method.

Author Contributions: Conceptualization, S.C. and C.G.L.; methodology, S.C.; validation, S.C., C.G.L. and J.H.Y.; formal analysis and investigation, S.C.; writing—original draft preparation, S.C.; writing—review and editing, S.C., C.G.L. and J.H.Y.; visualization, S.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research is supported by Jeollannam-do (2021 R&D supporting program operated by Jeonnam Technopark) and financially supported by the Ministry of Trade, Industry and Energy (MOTIE) and Korea Institute for Advancement of Technology (KIAT) under the research project: "National Innovation Cluster R&D program" (Grant number: P0016223).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data is available in a publicly accessible repository. The data used in this study are openly available from the Scientific Data portal in https://www.nature.com/articles/s41597-020-00721-w, accessed on 13 January 2022.

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

References

1. Dileep, G. A survey on smart grid technologies and applications. *Renew. Energy* **2020**, *146*, 2589–2625. [CrossRef]
2. Zhou, X.; Ma, Y.; Gao, Z.; Wang, H. Summary of smart metering and smart grid communication. In Proceedings of the IEEE International Conference on Mechatronics and Automation, Takamatsu, Japan, 6–9 August 2017. [CrossRef]
3. Ahmed, Z.G.; Ahmed, A.H.; Nabil, H.A. Dynamic Pricing; Different Schemes, Related Research Survey and Evaluation. In Proceedings of the International Renewable Energy Congress, Hammamet, Tunisia, 20–22 March 2018.
4. Nojavan, S.; Ajoulabadi, V.; Khalili, T.; Member, S.; Bidram, A.; Member, S. Optimal Power Flow Considering Time of Use and Real-Time Pricing Demand Response Programs. *arXiv* **2021**, arXiv:2102.07828.
5. Liu, X.; Golab, L.; Golab, W.; Ilyas, I.F. Benchmarking smart meter data analytics. In Proceedings of the EDBT 2015 18th International Conference on Extending Database Technology, Brussels, Belgium, 23–27 March 2015; pp. 385–396. [CrossRef]
6. Zhou, S.; Brown, M.A. Smart meter deployment in Europe: A comparative case study on the impacts of national policy schemes. *J. Clean. Prod.* **2017**, *144*, 22–32. [CrossRef]
7. Antonopoulosa, I.; Robu, V.; Couraud, B.; Kirli, D.; Norbu, S.; Kiprakis, A.; Flynn, D.; Elizondo-Gonzalez, S.; Wattame, S. Artificial intelligence and machine learning approaches to energy demand-side response: A systematic review. *Renew. Sustain. Energy Rev.* **2020**, *130*, 109899. [CrossRef]

8. Molchanov, V.; Linsen, L. Overcoming the curse of dimensionality when clustering multivariate volume data. In Proceedings of the VISIGRAPP 2018-Proceedings of the 13th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, Funchal, Madeira, Portugal, 27–29 January 2018. [CrossRef]
9. Meinrenken, C.J.; Rauschkolb, N.; Abrol, S.; Chakrabarty, T.; Decalf, V.C.; Hidey, C.; McKeown, K.; Mehmani, A.; Modi, V.; Culligan, P.J. MFRED, 10 second interval real and reactive power for groups of 390 US apartments of varying size and vintage. *Sci. Data* **2020**, *7*, 375. [CrossRef]
10. Saxena, A.; Prasad, M.; Gupta, A.; Bharill, N.; Pate, P.O.; Tiwari, A.; Joo, E.M.; Weiping, D.; Chin-Teng, L. A review of clustering techniques and developments. *Neurocomputing* **2017**, *267*, 664–681. [CrossRef]
11. Rajabi, A.; Li, L.; Zhang, J.; Zhu, J.; Ghavidel, S.; Ghadi, M.J. A review on clustering of residential electricity customers and its applications. In Proceedings of the 2017 20th International Conference on Electrical Machines and Systems (ICEMS), Sydney, Australia, 11–14 August 2017. [CrossRef]
12. Zhang, M.; Sun, S.; Cao, G.; Kong, X.; Zhao, X.; Zong, S. Load characteristics analysis based on improved k-means clustering algorithm. In Proceedings of the 2019 IEEE 2nd International Conference on Automation, Electronics and Electrical Engineering (AUTEEE), Shenyang, China, 22–24 November 2019; pp. 510–515. [CrossRef]
13. Panapakidis, I.; Asimopoulos, N.; Dagoumas, A.; Christoforidis, G.C. An improved fuzzy c-means algorithm for the implementation of demand side management measures. *Energies* **2017**, *10*, 1407. [CrossRef]
14. Luo, Z.; Hong, S.H.; Ding, Y.M. A data mining-driven incentive-based demand response scheme for a virtual power plant. Appl. *Energy* **2019**, *239*, 549–559. [CrossRef]
15. Stephen, H.; Colin, S.; Peter, G. Analysis and Clustering of Residential Customers Energy Behavioral Demand Using Smart Meter Data. *IEEE Trans. Smart Grid* **2016**, *7*, 135–144.
16. Bouveyron, C.; Girard, S.; Schmid, C. High-dimensional data clustering. *Comput. Stat. Data Anal.* **2007**, *52*, 502–519. [CrossRef]
17. Chen, T.; Qian, K.; Mutanen, A.; Schuller, B.; Jarventausta, P.; Su, W.W. Classification of electricity customer groups towards individualized price scheme design. In Proceedings of the 2017 North American Power Symposium (NAPS), Morgantown, WV, USA, 17–19 September 2017; pp. 4–7. [CrossRef]
18. Wang, J.; Wang, K.; Jia, R.; Chen, X. Research on Load Clustering Based on Singular Value Decomposition and K-means Clustering Algorithm. In Proceedings of the 2020 Asia Energy and Electrical Engineering Symposium (AEEES), Chengdu, China, 29–31 May 2020; pp. 831–835. [CrossRef]
19. Jiang, Z.; Lin, R.; Yang, F.; Wu, B. A fused load curve clustering algorithm based on wavelet transform. *IEEE Trans. Ind. Inform.* **2018**, *14*, 1856–1865. [CrossRef]
20. Vij, A.; Khandnor, P. Validity of internal cluster indices. In Proceedings of the 2016 International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS), Bengaluru, India, 6–8 October 2016; pp. 388–395. [CrossRef]
21. Li, F.; Tian, Y.; Wu, Y.; Liu, Y. A method of mining electricity consumption behaviour based on CC-DWT. In *IOP Conference Series: Earth and Environmental Science*; IOP Publishing: Bristol, UK, 2021; p. 012133. [CrossRef]
22. Aliyu, I.; Lim, C.G. Selection of optimal wavelet features for epileptic EEG signal classification with LSTM. *Neural Comput. Appl.* **2021**, 1–21. [CrossRef]
23. Phinyomark, A.; Limsakul, C.; Phukpattaranont, P. Application of wavelet analysis in EMG feature extraction for pattern classification. *Meas. Sci. Rev.* **2011**, *11*, 45–52. [CrossRef]
24. Rhif, M.; Ben Abbes, A.; Farah, I.R.; Martínez, B.; Sang, Y. Wavelet transform application for/in non-stationary time-series analysis: A review. *Appl. Sci.* **2019**, *9*, 1345. [CrossRef]
25. Cugliari, J.; Goude, Y.; Poggi, J.-M. Disaggregated Electricity Forecasting using Wavelet-Based Clustering of Individual Consumers. In Proceedings of the 2016 IEEE International Energy Conference (ENERGYCON), Leuven, Belgium, 4–8 April 2016. [CrossRef]
26. Gogtay, N.J.; Thatte, U.M. Principles of correlation analysis. *J. Assoc. Physicians India* **2017**, *65*, 78–81. [PubMed]
27. Das, S.; Rao, P.S.N. Principal Component Analysis based Compression Scheme for Power System Steady State Operational Data. In Proceedings of the ISGT2011-India, Kollam, India, 1–3 December 2011. [CrossRef]
28. 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]
29. Hartigan, J.A.; Wong, M.A. A K-means Clustering Algorithm. *J. R. Stat. Society. Ser. C (Appl. Stat.)* **1979**, *28*, 100–108.
30. Liu, F.; Deng, Y. Determine the Number of Unknown Targets in Open World Based on Elbow Method. *IEEE Trans. Fuzzy Systems.* **2021**, *29*, 986–995. [CrossRef]
31. Thinsungnoen, T.; Kaoungku, N.; Durongdumronchai, P.; Kerdprasop, K.; Kerdprasop, N. The Clustering Validity with Silhouette and Sum of Squared Errors. In Proceedings of the International Conference on Industrial Application Engineering, Kitakyushu, Japan, 28–31 March 2015; pp. 44–51. [CrossRef]
32. Bengfort, B.; Bilbro, R. Yellowbrick: Visualizing the Scikit-Learn Model Selection Process. *J. Open Source Softw.* **2019**, *4*, 1075. [CrossRef]
33. Chicco, G. Overview and performance assessment of the clustering methods for electrical load pattern grouping. *Energy* **2012**, *42*, 68–80. [CrossRef]
34. Ma, Z.; Yan, Y.; Li, K.; Nord, N. Building energy performance assessment using volatility change based symbolic transformation and hierarchical clustering. *Energy Build.* **2018**, *166*, 284–295. [CrossRef]

35. Vijaya, V.; Sharma, S.; Batra, N. Comparative Study of Single Linkage, Complete Linkage, and Ward Method of Agglomerative Clustering. In Proceedings of the of the International Conference on Machine Learning, Big Data, Cloud and Parallel Computing: Trends, Perspectives and Prospects (COMITC), Faridabad, India, 14–16 February 2019; pp. 568–573. [CrossRef]
36. Zhou, Z.; Yang, S.; Shao, Z. Household monthly electricity consumption pattern mining: A fuzzy clustering-based model and a case study. *J. Clean. Prod.* **2017**, *141*, 900–908. [CrossRef]
37. Vendramin, L.; Jaskowiak, P.A.; Campello, R.J.G.B. On the combination of relative clustering validity criteria. In Proceedings of the 25th International Conference on Scientific and Statistical Database Management (SSDBM), Baltimore, MA, USA, 29–31 July 2013; pp. 733–744. [CrossRef]