Building Energy Profile Clustering Based on Energy Consumption Patterns

Milad Afzalan

Thesis submitted to the faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Master of Science
In
Computer Science and Application

Hoda M. Eldardiry, Chair
Farrokh Jazizadeh, Co-chair
Edward A. Fox

June 8th, 2020
Blacksburg, Virginia

Keywords: Clustering, Unsupervised learning, Segmentation, Smart gird, Energy consumption.

Building Energy Profile Clustering Based on Energy Consumption Patterns

Milad Afzalan

ABSTRACT

With the widespread adoption of smart meters in buildings, an unprecedented amount of high-resolution energy data is released, which provides opportunities to understand building consumption patterns. Accordingly, research efforts have employed data analytics and machine learning methods for the segmentation of consumers based on their load profiles, which help utilities and energy providers for customized/personalized targeting for energy programs. However, building energy segmentation methodologies may present oversimplified representations of load shapes, which do not properly capture the realistic energy consumption patterns, in terms of temporal shapes and magnitude. In this thesis, we introduce a clustering technique that is capable of preserving both temporal patterns and total consumption of load shapes from customers' energy data. The proposed approach first overpopulates clusters as the initial stage to preserve the accuracy and merges the similar ones to reduce redundancy in the second stage by integrating time-series similarity techniques. For such a purpose, different time-series similarity measures based on Dynamic Time Warping (DTW) are employed. Furthermore, evaluations of different unsupervised clustering methods such as k-means, hierarchical clustering, fuzzy c-means, and self-organizing map were presented on building load shape portfolios, and their performance were quantitatively and qualitatively compared. The evaluation was carried out on real energy data of ~250 households. The comparative assessment (both qualitatively and quantitatively) demonstrated the applicability of the proposed approach compared to benchmark techniques for power time-series clustering of household load shapes. The contribution of this thesis is to: (1) present a comparative assessment of clustering techniques on household electricity load shapes and highlighting the inadequacy of conventional validation indices for choosing the cluster number and (2) propose a two-stage clustering approach to improve the representation of temporal patterns and magnitude of household load shapes.

Building Energy Profile Clustering Based on Energy Consumption Patterns

Milad Afzalan

GENERAL AUDIENCE ABSTRACT

With the unprecedented amount of data collected by smart meters, we have opportunities to systematically analyze the energy consumption patterns of households. Specifically, through using data analytics methods, one could cluster a large number of energy patterns (collected on a daily basis) into a number of representative groups, which could reveal actionable patterns for electric utilities for energy planning. However, commonly used clustering approaches may not properly show the variation of energy patterns or energy volume of customers at a neighborhood scale. Therefore, in this thesis, we introduced a clustering approach to improve the cluster representation by preserving the temporal shapes and energy volume of daily profiles (i.e., the energy data of a household collected during 1 day). In the first part of the study, we evaluated several well-known clustering techniques and validation indices in the literature and showed that they do not necessarily work well for this domain-specific problem. As a result, in the second part, we introduced a two-stage clustering technique to extract the typical energy consumption patterns of households. Different visualization and quantified metrics are shown for the comparison and applicability of the methods. A case-study on several datasets comprising more than 250 households was considered for evaluation. The findings show that datasets with more than thousands of observations can be clustered into 10-50 groups through the introduced two-stage approach, while reasonably maintaining the energy patterns and energy volume of individual profiles.

ACKNOWLEDGMENT

I would like to express my sincere gratitude to my advisor, Dr. Hoda Eldardiry, for all her support, guidance, and encouragement. I would like to thank my committee members, Dr. Farrokh Jazizdeh (co-advisor), and Dr. Edward Fox for their constructive feedback and recommendations.

I am grateful to my family, for their invaluable love and support. Your encouragement has always been my greatest motivation to pursue my dreams.

Table of contents

List of figures

List of tables

Chapter 1: Introduction and motivation

Smart meters, which are being deployed at the nationwide scale, produce a vast amount of electricity data from residential households. This amount of high-resolution electricity data, which is collected typically in sub-hourly resolution from many households, provides opportunities for mining the energy consumption patterns. Specifically, electricity load shapes, which are time-series of energy data collected at the span of one day, could be characterized by clustering techniques to provide an overview of typical energy consumption patterns of households. To this end, segmentation, which is the task of clustering energy load shapes from a large number of households, could reveal an actionable pattern for utilities to target households for customized energy programs that are tailored to household usage style. Examples of energy programs include demand response (identifying households with high electricity demand at peak time and encouraging them to reduce load) or engagement of households for adopting renewable resources such as solar panels.

During recent years, the application of segmentation on household electricity load shapes has received increasing attention. The objective of data-driven segmentation is to group customers based on similar energy behavior patterns and energy volume into representative clusters. Considering the importance of this problem, in this thesis, we first present a comparative assessment of clustering techniques on household electricity load shapes. Furthermore, we introduce a two-stage clustering technique to improve benchmark techniques through an improved representation of load shapes. A case-study on multiple datasets collected from residential buildings is presented for comparison and to show the applicability of the approach.

1.1. Problem statement and research gaps

The research problem addressed in this thesis is to recover the distinct temporal shapes and power magnitude of household energy profiles through clustering. Due to the highly varied patterns of energy consumption across households on a daily basis, the task of segmentation can become challenging, especially when the dataset gets large. As an example, Figure *1-1* shows the clustering results of ~8000 load shapes into 5 clusters using K-means, which is commonly applied in the literature for datasets of similar size. In Figure *1-1*(a), each subplot shows the temporal shape of each cluster, and the vertical line is the power magnitude (in kW). The red line is the centroid of

each cluster, which is obtained by averaging all daily profiles (in the order of hundreds or thousands) associated with the cluster. In Figure *1-1*(b), several examples of daily profiles associated with clusters 1 and 4 are shown. In many prior efforts, the segmentation results were presented similar to what is shown in Figure *1-1*(a), without further evaluation of cluster quality. In other words, the results of clustering were taken as self-evident, and no further investigations were carried out to show the representativeness of each cluster to their associated members. For example, as can be seen in Figure *1-1*(b), examples of load shapes for cluster 1 are retrieved that do not resemble the shape of the cluster 1 centroid. Similarly, a number of load shapes in cluster 4 are presented (shown in part (b)) that do not share similarity with its centroid (part (a)). Considering that examples of load shapes in Figure *1-1*(b) may have a considerable density in the entire dataset, it is important to form new clusters for them due to their distinct energy behavior. While one solution is to increase the number of clusters to allow for better representation of load shapes, one should keep in mind that this comes at the cost of obtaining a large number of clusters with high similarity, that contradicts the objective of segmentation.

Based on this discussion, variations in temporal patterns of load shapes and their power magnitude make the segmentation task challenging. Specifically, this problem becomes more important as the scope of data gets larger, and a higher number of households and historical days would be considered. Therefore, it is important to perform efficient clustering on such datasets to allow for better representation of typical energy consumption patterns while maintaining the interpretability of the clustering task.

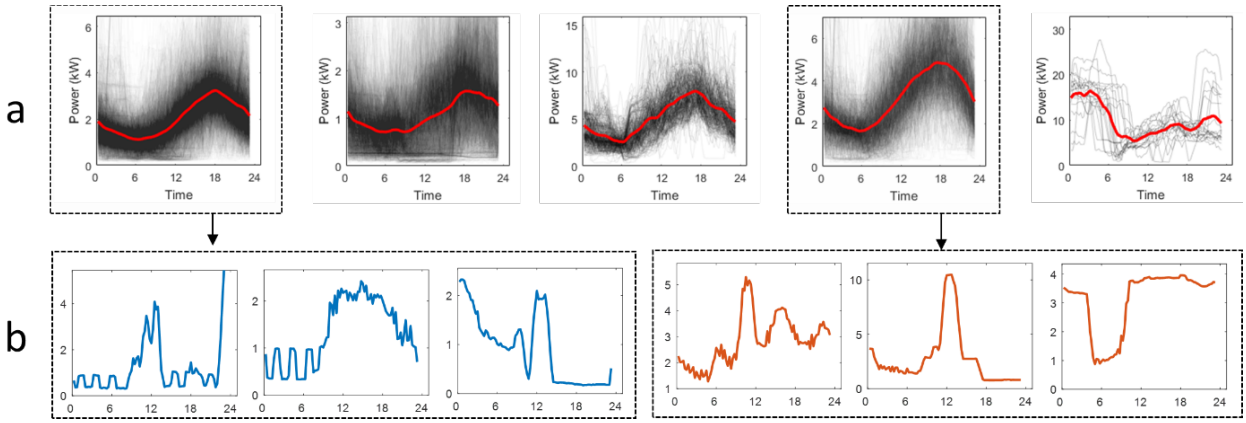


Figure 1-1. Segmentation of ~8000 load shapes using K-means: (a) five clusters representing the entire dataset, (b) examples of individual daily profiles for clusters 1 and 4 that are wrongly classified.

1.2. Contributions

In this thesis, we (1) compared the clustering results based on the two dimensions of (i) different clustering techniques and (ii) different cluster validation index (CVI). Furthermore, it was shown that common CVIs for these specific problems have a shortcoming for the realistic representation of electricity load shapes, (2) we introduced a two-stage clustering approach to investigate this problem. The proposed approach in the first stage overpopulates a larger number of clusters to improve the accuracy and it then merges similar clusters by integrating several time-series modeling techniques in the second stage.

1.3. Thesis structure

The rest of the thesis is structured as follows: In chapter 2, the literature review is presented. Chapter 3 presents the case-study datasets. Chapter 4 presents the comparative performance of clustering techniques on electricity load shapes and further discusses the CVI results. Chapter 5 introduces the two-stage clustering technique and presents the findings and evaluations. Chapter 6 concludes the thesis and outlines future directions.

Chapter 2: Literature review

The roll-out of smart metering infrastructures [1] has provided opportunities to capture the consumption information of residential buildings with higher resolution (e.g., every 15-minute or per-minute). Accordingly, recent attempts (e.g., [2-18]) have focused on characterizing the usage behavior into certain similar patterns, with the aim of facilitating the selection of consumers for improved targeting for different customized energy programs. In this context, energy programs are attributed to those that focus on improving demand-side management (DSM) (i.e., reshaping the energy demand profiles) and demand response (DR) (i.e., shifting energy loads at peak time) [12, 19], the adoption of renewable energy resources such as solar panels [20], and improving energy efficiency attributes (i.e., replacing appliances with more energy-efficient ones or using interventions for energy behavioral change) [21]. As a common attribute for the aforementioned applications, different studies have used temporal consumption data of homes for learning and characterizing the usage behavior. Accordingly, the objective was to introduce user segmentation methodologies for the collective targeting of users for specific programs that are tailored to their consumption patterns. Depending on the scope of studies, household energy consumption behavior can be analyzed in user, time, and spatial dimensions [22]. As an example of user and time dimensions, Figure *2-1* shows a sample of segmentation results obtained by K-means clustering on load profiles of ~11000 buildings [23]. As can be seen, load shapes can be categorized into different groups, which in turn could enable characterizing the consumption patterns of homes based on the occurrence of load shapes over multiple days.

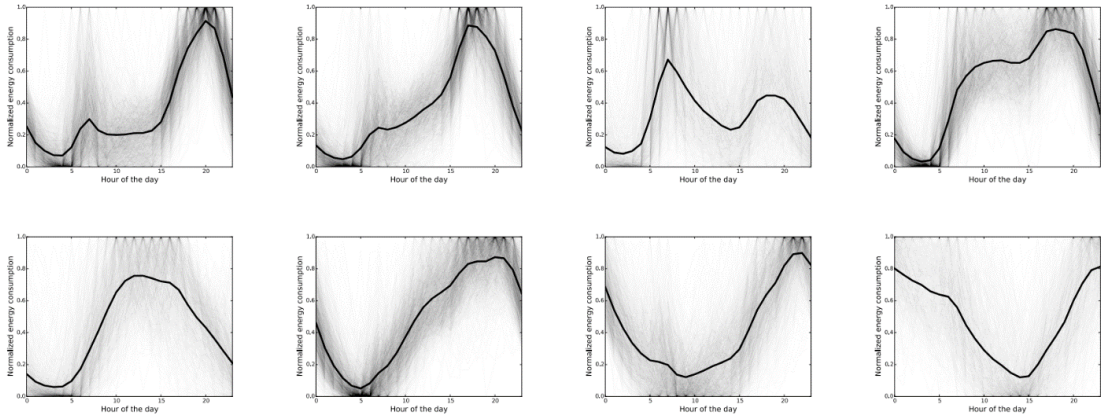


Figure 2-1. Segmentation results based on clusters of load profiles for ~11000 buildings. Taken from reference [23].

Several notable preliminary works in the user segmentation field can be found in [3, 6, 8, 10]. The applications of different clustering methods such as k-means [4, 8, 24], hierarchical [4], self-organizing maps (SOM) [10, 25], and expectation maximization (EM) [26] have been investigated. Different validation metrics including clustering dispersion indicator (CDI) and Davies-Bouldin Indicator (BDI) have been evaluated [4, 8], and the potential implications of consumer classification for load forecasting [27] and determination of tariffs [28] were explored. These preliminary studies had a focus on rather a short amount of data, and the complexity that arises in the analysis of large-scale data (orders of hundreds or thousands of homes over months of data) has not been addressed. To this end, segmentation methods have looked into scalable and cost-efficient solutions to address the complexity of large-scale problems [29]. Kwac et al. [12] proposed a segmentation methodology using hierarchical and adaptive k-means clustering for household segmentation using hourly smart meter data. The output of their work was a sample of load shape which could be employed for targeting homes for DR. In addition, they used the notion of entropy on the distribution of load shapes for each home to determine the consistency of usage across different days. In this context, consistency has been defined as a measure for variability or stability of load shapes over days. The same authors later leveraged the results in [12] and used the distribution of load shapes for each home to make inference on the lifestyle patterns (i.e., probability distribution of representative features for home energy consumption) [30]. To improve their segmentation results, they used the earth mover distance (EMD) metric to consider the temporal location of values for calculating the distance between shapes. A later study [31] showed that the segmentation methods used in [12] can lead to a large population of load shapes, which further makes it difficult to classify consumers into representative groups. A similarity measure based on the Dynamic Time Warping (DTW) [32] was employed to consider the optimal alignment between consumption style patterns, which showed that a 50% reduction in the number of clusters can be achieved with improvement in prediction accuracy. The application of user segmentation for renewable integration has been further investigated in [20], in which the amounts for solar PV and battery capacity for each group was determined.

Considering the discussion, we have summarized the characteristics of major studies that have focused on the segmentation task for a specific objective/application in Table *2-1*.

Table 2-1. Summary of major user (home) segmentation studies and their characteristics

Ref	Method/metrics used	Primary application	Dataset location	Data type
[12]	Hierarchical and k-means	Targeting for DSM	CA, USA	Aggregate
[30]	Hierarchical and k-means Using EMD metric	Lifestyle segmentation	CA, USA	Aggregate
[2]	SOM	Examined the impact of load shapes on homes' characteristics	Ireland	Aggregate
[31]	Using EMD metric	Reducing the number of representative load shapes for segmentation	CA, USA	Aggregate
[20]	Hierarchical and k-means	PV and battery storage adoption	TX, USA	Aggregate
[7]	Clusterwise regression and k-means	Comparison of predictive accuracy and cluster stability for segmentation	NY, USA	Aggregate
[15]	K-means	Determined correlation of clustering results and homes' survey data	Austin, TX	Aggregate
[33]	Hierarchical clustering with EMD metric	Improved forecasting performance	Austin, TX	Aggregate
[34]	K-means	Estimated peak reduction amount from AC units based on the temporal shape and peak magnitude values of clustered load shapes	Australia	Appliance –level (AC)
[35]	Piecewise aggregate approximation/ spectral clustering	Dimensionality reduction of load shapes while preserving accuracy	Shanghai, China	Appliance –level (AC)

Recent efforts in household electricity segmentation have focused on addressing specific problems. In a class of studies, the use of big data, and the efficiency of clustering at the city-scale

have been investigated [23, 29]. Therefore, due to the increasing size of datasets, the applicability of dimensionality reduction techniques such as Principal Component Analysis (PCA) and Symbolic Aggregate Approximation (SAX) [36], in addition to feature extraction on the attributes of interests on load shapes (such as peak distributions or key time frames) have been investigated [37, 38]. Furthermore, to mitigate the impact of noisy and unequal time-series with incomplete information in real-world scenarios, the model-based approach which accounts for phase shift and time lag has been proposed [18]. In some recent attempts, investigation of segmentation on decomposed data, including Air Conditioning (AC) has been carried out for DR applications [39, 40].

As a core component of household electricity segmentation and the potential applications that it can offer, the clustering output (both in terms of cluster quality and the number of clusters) is assumed to reasonably represent the energy behavior patterns of the entire customer base. Such an assumption has been taken self-evident without further investigation [15, 41], studied through common cluster validation indices (CVI) [2, 4, 36] or justified through visual inspection [23]. Among such approaches, using typical CVI may be considered to better ensure the quality of segmentation at first sight. Examples of typical CVI for clustering include Bayesian Information Criterion (BIC), Silhouette index, and Davies–Bouldin index (DBI). However, as outlined in [23], such generic statistical indicators for model selection will not necessarily work for the electricity segmentation task. Furthermore, proper metrics, that indeed capture the representativeness of clusters with their associated profiles, are mainly ignored in the evaluation process [20].

As the trend in the literature shows, the household electricity segmentation works have mainly overlooked the importance of recovering typical load shapes with distinct temporal patterns. Therefore, in this work, we focus on the task of identifying load shapes with distinct temporal peaks and energy magnitude. A two-stage clustering approach has been proposed. In the first stage, the focus is on recovering distinct load shapes. In the second stage, we employ efficient merging techniques to identify correlated clusters and reduce the cluster library size.

Chapter 3: Case-study dataset

3.1. Dataset selection

To present the findings and evaluations, we needed to have a dataset collected from multiple buildings and subsequent days. In this way, we could have assumed to have enough variations in energy consumption patterns for the evaluation. In this work, we used the data from the Pecan Street Project [42], which is an ongoing campaign in energy-efficiency initiative through equipping residential buildings with metering devices. Here, we used a subset of a dataset that was collected from residential buildings in Austin, TX and Boulder, CO during July and August 2015. The resolution of data was 15-minute per sample, therefore each daily profile contains 96 data points. Three datasets were considered as follows:

- Dataset 1: Location: Austin, TX, number of households: 129, total number of daily profiles: 7535, number of samples per daily profile: 96.
- Dataset 2: Location: Austin, TX, number of households: 100, total number of daily profiles: 5676, number of samples per daily profile: 96.
- Dataset 3: Location: Boulder, CO, number of households: 31, total number of daily profiles: 1790, number of samples per daily profile: 96.

Table *3-1* presents the characteristics of the datasets.

Table 3-1. Characteristics of the dataset.

Dataset	Location	Number of households	Duration	# of daily profiles	# of samples per profile
Dataset 1	Austin, TX	129	60 days	7535	96
Dataset 2	Austin, TX	100	60 days	5676	96
Dataset 3	Boulder, CO	31	60 days	1790	96

Figure *3-1* shows the distribution of energy among all profiles for different datasets. The median energy consumption for dataset 1, dataset 2, and dataset 3 is 44 kWh, 42kWh, and 17kWh.

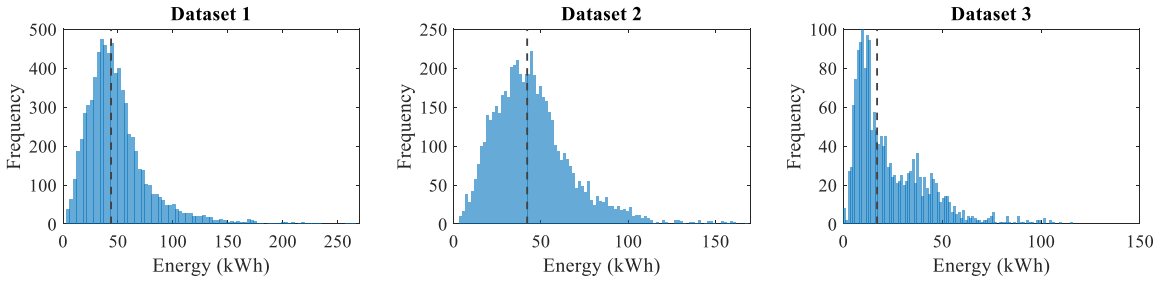


Figure 3-1. Distribution of energy among all profiles for different datasets. The dashed line shows the median.

3.2. Dataset cleaning and processing

To prepare the dataset for the analysis, it was stored in a matrix structure with the form $n \times m$, in which n is the number of daily profiles, and m is the number of samples appended by the unique ID of profiles. n values are the same as what was mentioned in Section 3.1 for each dataset, and m is set as 98, in which the first 96 elements are a data sample (collected every 15-minute at the span of 24-hour) and the last two elements are unique identifiers of the profile with the household ID and day of the year. Each profile is annotated with its unique identifier for the sake of information retrieval.

Given the high-resolution of data, the impact of noise could adversely impact the clustering performance. To reduce the impact of noise and artifact, a moving average filtering was performed on the data. A window of $w = 4$ data points, as 1 hour, was considered. Figure *3-2* shows examples of daily profiles and their filtered version with a moving average.

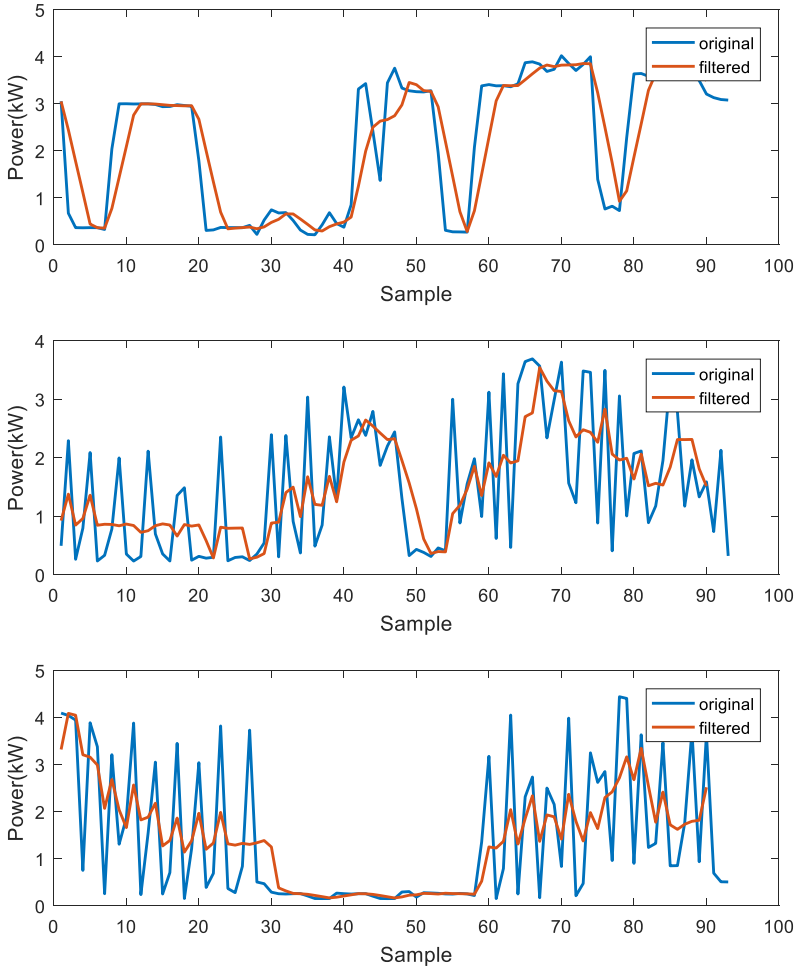


Figure 3-2. Examples of original daily profiles and filtered profiles with a moving average.

Chapter 4: Comparative assessment of clustering techniques on electricity load shapes

In this chapter, we present the comparative assessment of conventional clustering techniques on electricity load shapes. Specifically, we investigated K-means, hierarchical clustering, and self-organizing map (SOM) as three different methods for evaluation. The selection of these methods is based on their applicability to time-series clustering. Comparison through cluster validation indices (CVI) is presented.

4.1. Clustering algorithms

4.1.1. K-means

K-means is one of the most widely used statistical clustering approaches. K-means, as a distance measure-based approach, minimizes a sum of mean squared error such that:

$$SSE = \sum_{k=1}^{K} \sum_{x \in C_k} \|x - C_k\|^2 \quad (1)$$

in which K is the total number of clusters, and C_k is the centroid of cluster k. Basically, K-means takes K clusters as the predetermined value, and randomly assigns K data points as the initial clusters. Thereafter, the following process is repeated: (1) assign each data point to the closest cluster centroid, (2) recalculate each cluster centroid based on the data points that are associated with it. This process is repeated until the convergence criterion is met. The convergence can be defined when the assignments of data point to clusters do not change over iterations or through selecting a maximum number of iterations. K-means, by default, employs the widely used Euclidean distance for measuring the distance between observations. However, alternate distance measures are used in similar techniques such as K-medoids.

4.1.2. Fuzzy c-means

Fuzzy c-means approach is similar to the K-means clustering family. However, in fuzzy c-means, each data point has an extent of membership to each cluster. The degree of membership for each data point is subject to the following constraint:

$$\sum_{k=1}^{K} \mu_{ik} = 1 \quad (2)$$

in which μ_{ik} is the degree of membership for observation i to the cluster k. Similar to k-means, fuzzy c-means minimizes a sum of squared error such that:

$$\sum_{i=1}^{N}\sum_{k=1}^{K}\mu_{ik}^{m}\|x_i - C_k\|^2 \quad (3)$$

in which m adjusts the level of fuzziness. Each member is associated with the cluster that has the highest degree of membership value. The degree of membership values is also updated at each iteration.

4.1.3. Hierarchical clustering

Hierarchical clustering creates a dendrogram either through an agglomerative (bottom-up) or divisive (top-down) approach. In the former, each data point is initially treated as one cluster and then they are iteratively merged until one root cluster is obtained. In the latter, one single cluster is initially assumed and it is split iteratively to create the dendrogram structure. The dendrogram can be cut at any point of the tree structure to form the final clusters. As shown in the power domain clustering literature, the agglomerative approach is considered as the favorite solution [36]. In order to decide the similarity of clusters for hierarchical merging, a linkage criterion is required. The linkage criterion is measured over a similarity matrix, in which the distance between each observation i and j is presented. Main linkage criteria include single, complete, centroid, average, and Ward. In single linkage, the merging is based on the closest pair of data points belonging to different clusters. In the complete linkage, the merging is based on the dissimilarity of farthest data points belonging to different clusters. Since single and complete linkage relies on individual instances and therefore they are sensitive to individual elements, alternatives like average or centroid linkage can be used. The average linkage looks at the average dissimilarity of all pairs of data points belonging to different clusters while the centroid linkage measures the dissimilarity between centroids of different clusters. The Ward criterion uses an objective function to minimize the total sum of squared error by merging a pair of clusters. In this work, the Ward criterion is used as the linkage criteria.

4.1.4. Self-Organizing Map (SOM)

SOM is an unsupervised neural network-based clustering approach. In SOM, the input space is mapped to a reduced output space, in which nodes are sorted in a 2-D grid. Through different iterations, nodes on the grid converge to the areas with higher density, which reflect the underlying clusters in the dataset. The iteration in SOM includes: (1) creating the grid's node network on the

data space ($b_1 \times b_2$ map), (2) selecting a data point and finding the node that is closest to the data point. The node is referred to as the best matching unit (BMU), (3) move the BMU closer to the data point based on a learning rate, (4) move the neighbors of the BMU closer to the data point based on a BMU radius, while having the farthest one moving less, (5) update the learning rate and BMU radius with lower value and repeat the process until the grid's node network is stable. The selection of BMU is based on the following equation:

$$b = argmin_i \|x - n_i\| \tag{4}$$

in which n_i is the representation of the node i, which has the same dimension as the data points x. The content of each node vector n_i is updated in different iterations as:

$$n_i(t+1) = n_i(t) + \alpha(t)\theta_{bi}\big(x(t) - n_i(t)\big) \tag{5}$$

in which $\alpha(t)$ is the learning rate at time t, and θ_{bi} neighborhood kernel for the unit b. The neighbor nodes of the BMU are activated based on:

$$\theta_{bi}(t) = \exp\left(-\frac{\|r_b - r_i\|^2}{2\sigma^2(t)}\right) \tag{6}$$

Here, r denotes the coordinates of the units, and therefore, the numerator denotes the radius around BMU. σ is the neighborhood radius. Both α and σ values decrease with iterations.

4.2. Cluster validation indices

External validation, as a measure of the goodness of clustering, is an important task for evaluation. Since many unsupervised problems, in nature, do not have access to labeled data, a form of internal validation should be carried out. In internal validation, the goodness of clustering is only measured based on the information of data. Specific criteria for internal validation include compactness and separation [43]. Compactness is a measure of similarity among observations within each group (i.e., intra-cluster compactness), while separation is a measure of distinctness among different groups (i.e., inter-cluster separation). Here, several cluster validation indices (CVI) used in this work are presented:

- Davies-Bouldin index: For Davies-Bouldin index (DBI) [44], for each cluster $\boldsymbol{C_i}$, the similarity between a given cluster and all others is measured, and the maximum value is assigned to $\boldsymbol{C_i}$. Through averaging the similarity for all clusters, the DBI is measured. Mathematically:

$$DBI = \frac{1}{K}\sum\nolimits_i max_{j,j\neq i}\left\{\frac{\left[\frac{1}{\|C_i\|}\sum_{x\in C_i} d(x,\mu_i) + \frac{1}{\|C_j\|}\sum_{x\in C_j} d\left(x,\mu_j\right)\right]}{d\left(\mu_i,\mu_j\right)}\right\} \quad (7)$$

in which K is the number of clusters, $\|C_i\|$ is the number of observations in cluster i, μ_i is the centroid of cluster C_i, and d is the distance function.

As the definition of DBI implies, a lower value indicates better clustering.

- ***Sillhouette index:*** The silhouette index (SIL) [45] looks at both the pairwise distance of between and within cluster distances. Specifically:

$$SIL = \frac{1}{K}\sum\nolimits_i \left\{\frac{1}{\|C_i\|}\sum\nolimits_{x\in C_i}\frac{b(i)-a(i)}{\max\left(b(i),a(i)\right)}\right\} \quad (8)$$

in which

$$a(x) = \frac{1}{\|C_i\|-1}\sum\nolimits_{j\in C_i, i\neq j} d(i,j), \quad (9)$$

$$b(x) = min_{k\neq i}\frac{1}{\|C_i\|}\sum\nolimits_{j\in C_k} d(i,j)$$

Sillhouette index aims at maximizing its value for better clustering.

- ***Calinski-Harabasz index:*** The Calinski-Harabasz index (CHI) [46] looks at the average of between and within cluster sum of squares. Specifically:

$$CHI = \frac{\sum_i \|C_i\| * d^2(\mu_i,\mu_j)/(K-1)}{\sum_i \sum_{x\in C_i} * d^2\left(x,\mu_j\right)/(N-K)} \quad (10)$$

in which N is the total number of observations in the dataset.

CHI looks at maximizing its value for better clustering.

Within cluster sum of square error (WCSSE): The within sum of square error (WCSS) looks at minimizing the within-cluster sum of squared error as follows:

$$WCSSE = \sum\nolimits_i \sum\nolimits_{x\in C_i} \|x-\mu_i\|^2 \quad (11)$$

A lower value for WCSSE is desired. However, for selecting the proper number of clusters over a range of K, WCSSE uses the elbow criterion, in which the reduction in WCSSE after a specific value of K gets negligible compared to previous values.

4.3. Results and discussion

4.3.1. CVI comparison

In this section, the results of different clustering techniques, based on different CVI values are presented and compared. To evaluate the proper number of clusters based on a CVI, a range of different K's are considered and the CVI value is measured. Thereafter, based on the criteria described in the previous section, the maximum, minimum, or a proper value interpreted from an elbow curve would be selected as the final K. We considered the range $K = \{5, 10,15, 20, \dots ,120\}$ to have a reasonable estimation of low to high number of clusters for comparison. Since prior efforts mainly investigated the number of clusters between 5 to 10, we considered 1 cluster increment for that range (i.e., {5,6,7,8,9,10}), and from 10 afterward, we considered 5 cluster increments till $K(end) = 120$ (i.e., {10,15,20,…,120}). For each value of K, clustering is performed and CVIs are measured. Figure *4-1*, Figure *4-2*, and Figure *4-3* present the CVIs for Dataset 1, Dataset 2, and Dataset 3, respectively. Each subplot represents one of the CVIs. The initial examinations showed that the trend of increase/decrease for all 4 CVIs (each subplot) is consistent across all datasets. Therefore, we present the interpretation for Dataset 1 (Figure *4-1*), while the same interpretation can be made for Figure *4-2* and Figure *4-3* as well.

DBI metric uses a min-rule for selecting the K number. As the results show, for all methods, a low value of $K = 5$ leads to low DBI value. It must be noted that DBI values for fuzzy c-means showed to be very high. Therefore, we did not present its values in the first subplot to avoid masking the changes reflected by other methods with lower values. The SIL metric uses a max-rule for selecting the K. Similar to the previous case, the lowest value of $K = 5$ governs these criteria. The CHI metric uses a max-rule for selecting the K. A value of $K = 5$ is also selected for this metric. However, for WCSS, the elbow curve shows a proper value of $K = 30$. Regarding the techniques, for DBI and SIL, the K-means shows marginal improvement over SOM and hierarchical clustering. For the CH, SOM and K-means show a better performance. For the WCSS, hierarchical clustering has the lowest values and seems to have a better performance. However, to present the performance, empirical observation on the clustering results needs to be carried out.

Considering the above discussion and accounting for all factors together, we have shown multiple scenarios for each dataset that are empirically investigated in the next section. Table 4-1 presents different techniques and K for the empirical investigation in the next section.

Table 4-1. Techniques and K values for different datasets.

Dataset	Technique	*K*
1	K-means	5
	HC, SOM	30
2	K-means	8
	HC, SOM	25
3	K-means	6
	HC, SOM	25

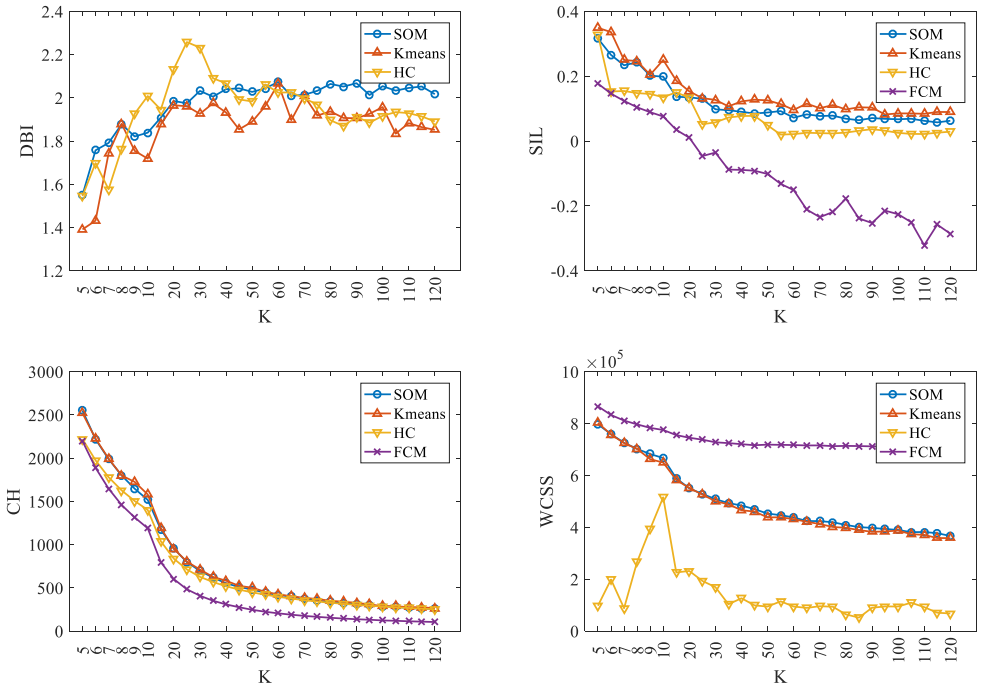


Figure 4-1. CVIs for different techniques for dataset 1.

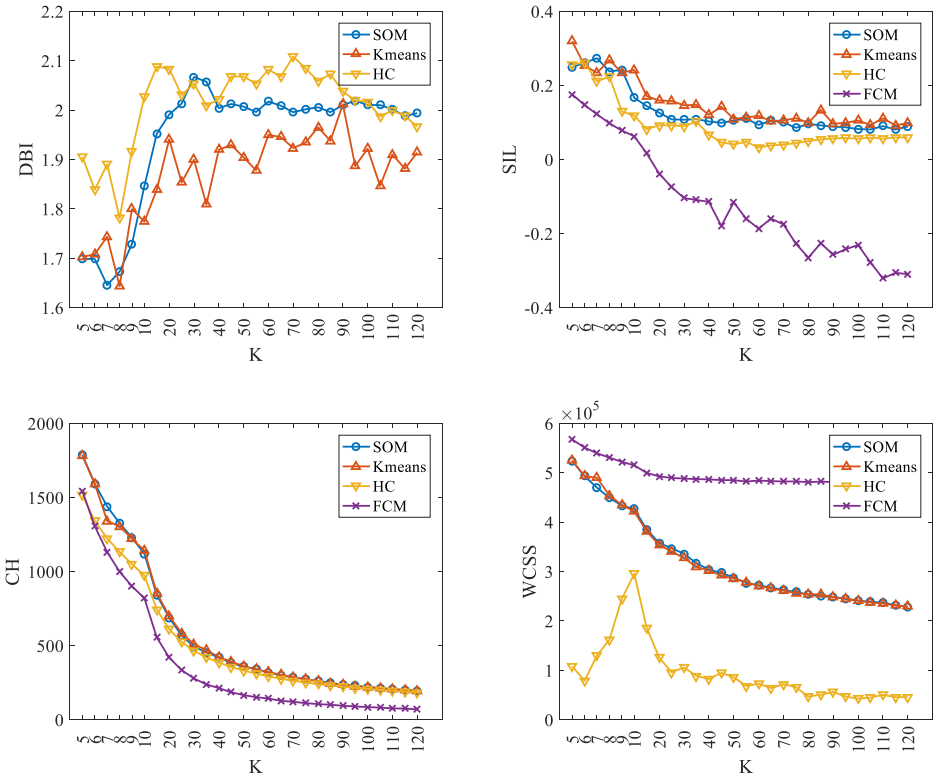


Figure 4-2. CVIs for different techniques for dataset 2.

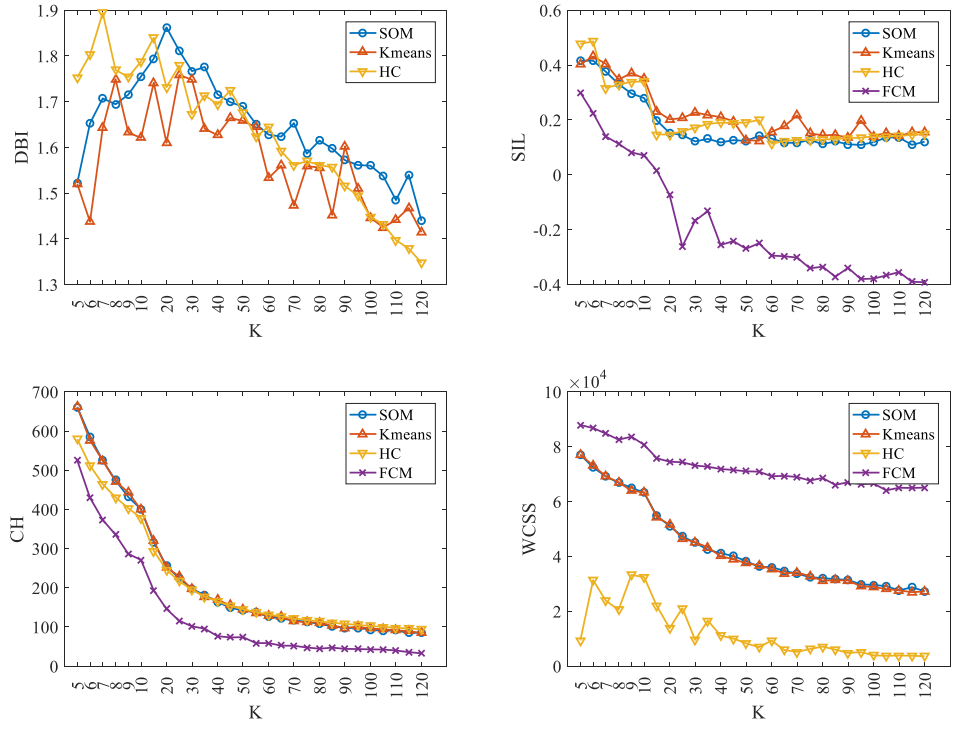


Figure 4-3. CVIs for different techniques for dataset 3.

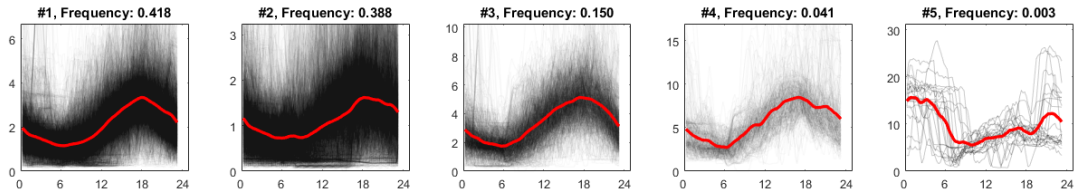
4.3.2. Empirical demonstration

To demonstrate the output of different scenarios described in Table *4-1*, Figure *4-4* to Figure *4-6* present the clusters for dataset 1, dataset 2, and dataset 3, respectively. Each subplot is one cluster, and the red curve is the cluster centroid, averaged over all observations associated with a cluster. The frequency value above each subplot shows the density of the cluster on the entire dataset. In Figure *4-4*(a), with $K = 5$, four out of five clusters have peak demand around 18:00, but with different peak magnitudes. Cluster 5 seems to be an outlier, due to the considerable high peak usage in addition to its very low frequency (<1% of the data). In Figure *4-4*(b) and Figure *4-4*(c), with a higher population of clusters, more distinct energy patterns are revealed, while their presence has been masked by selecting a lower number in Figure *4-4*(a). Examples include clusters 14, 20, and 22 in Figure *4-4*(b), and clusters 8, 18, and 22 in Figure *4-4*(c). Furthermore, a comparison between HC and SOM with an equal number of clusters (Figure *4-4*(b) and Figure *4-4*(c)) shows that HC gives more emphasis on identifying outlier clusters with very low frequency (4 clusters with a frequency of less than 0.5%) while SOM form clusters with higher density. As the results in Figure *4-4* show, although CVI selects K=5, using such a low value will result in a very coarse-level representation of load shapes. Specifically, the cluster centroids, which altogether are supposed to encompass the community behavior, might not reflect the temporal shape/power magnitude of their associated members.

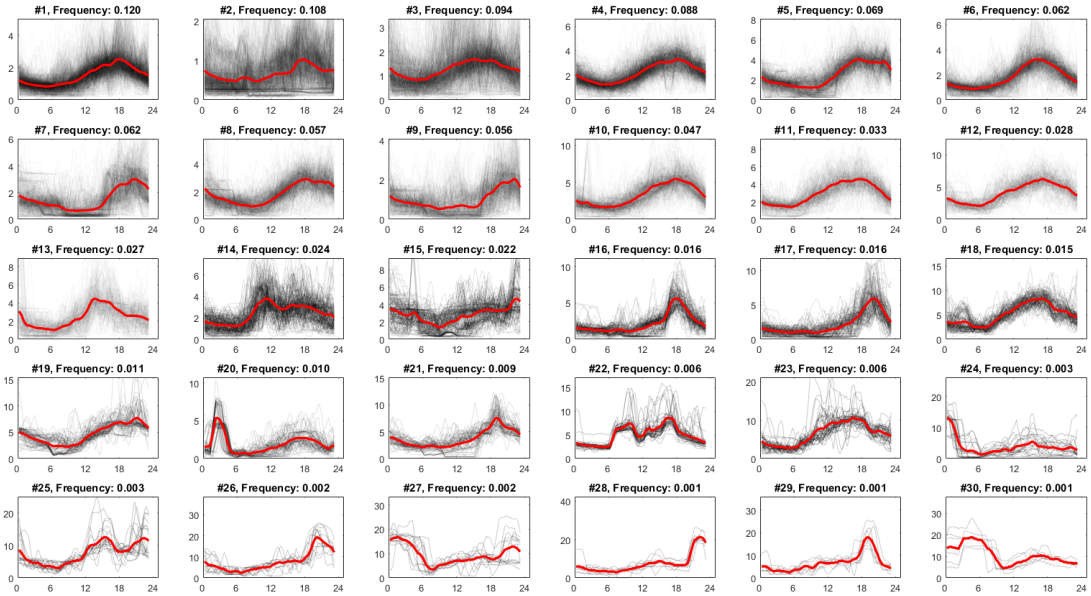
Similar to what was presented for Figure *4-4*, in Figure *4-5* and Figure *4-6* we observed a number of patterns with higher number of clusters (parts (b) and (c)), which were not recovered as a distinct group while selecting a lower number of clusters (part (a)). Examples include clusters 11 and 16 (Figure *4-5*(b)), and clusters 8 and 12 (Figure *4-5*(c)), in addition to cluster 14 (Figure *4-6*(b)) and cluster 5 (Figure *4-6*(c)). From the application perspective, an appropriate segmentation approach needs to recover distinct energy patterns from the clustering stage. For example, in Figure *4-6*(c)), identifying cluster 5 is important, since it is applicable for benefiting from installing solar panels (see the sharp peak demand at noon, the same time as solar generation). However, with a mere focus on CVI values, such groups may not be identified (part (a) in Figure *4-4* to Figure *4-6*).

To summarize the discussion, there exist some variations in CVIs, and using different metrics result in selecting both different K values and different techniques. Furthermore, generic CVI might not work well for this domain-specific problem, because several groups of energy patterns with distinct load shapes/magnitude that could be of interest to energy planners/utilities may not

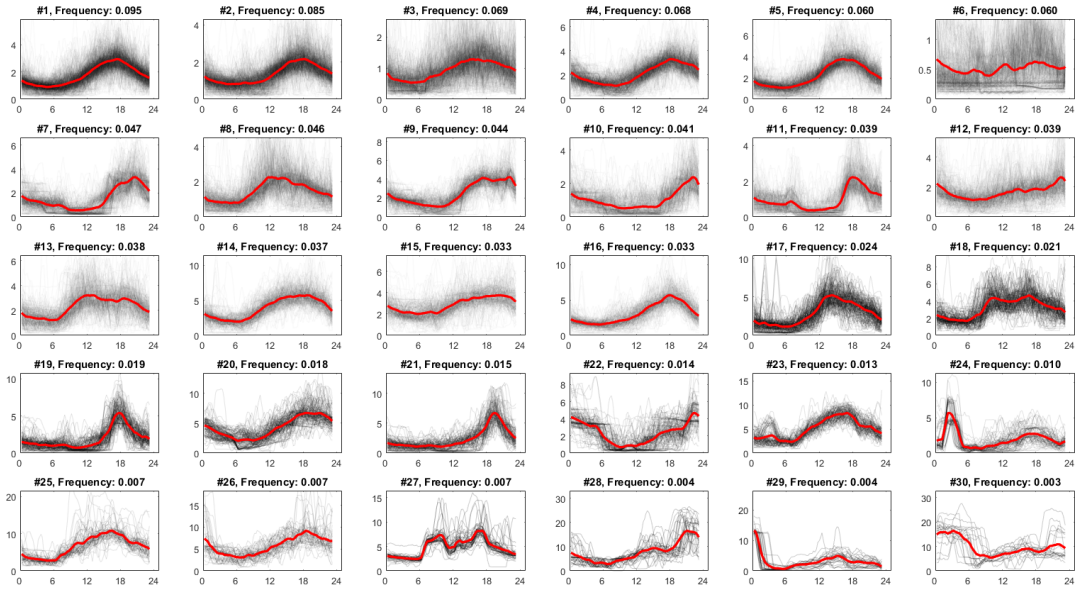
be recovered without selecting a higher number of clusters. Therefore, the next chapter of the thesis investigates this problem.



(a)

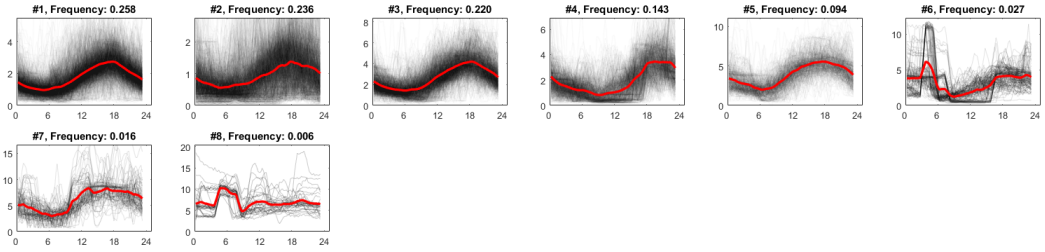


(b)

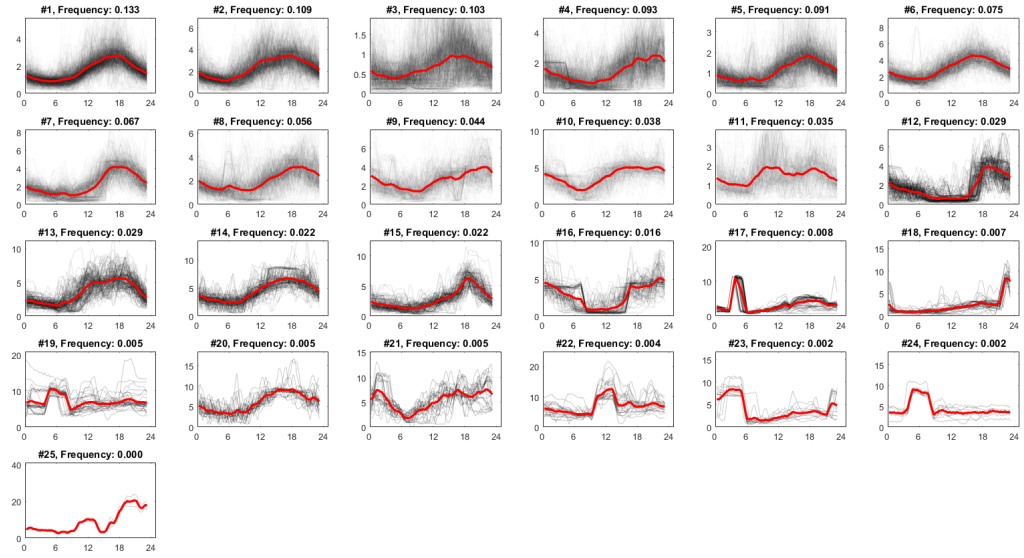


(c)

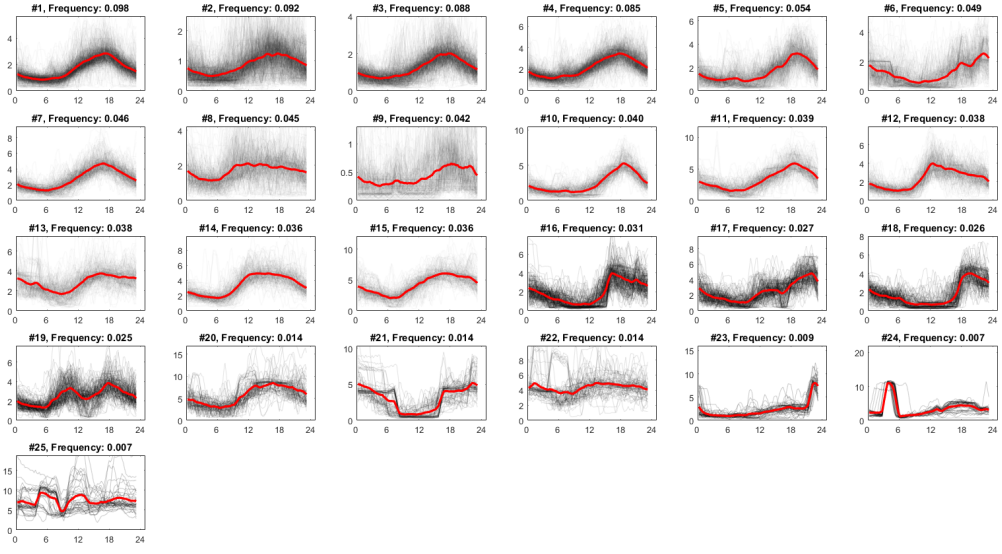
Figure 4-4. Cluster for dataset 1: (a) K-means, K=10, (b) HC, K=30, (c) SOM, K=30. The vertical axis is 'Power (kW)' and the horizontal axis is 'Time (hr)'.



(a)

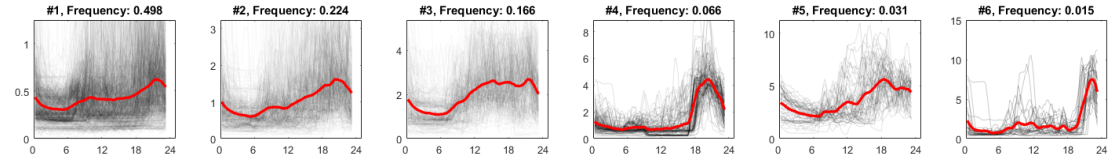


(b)

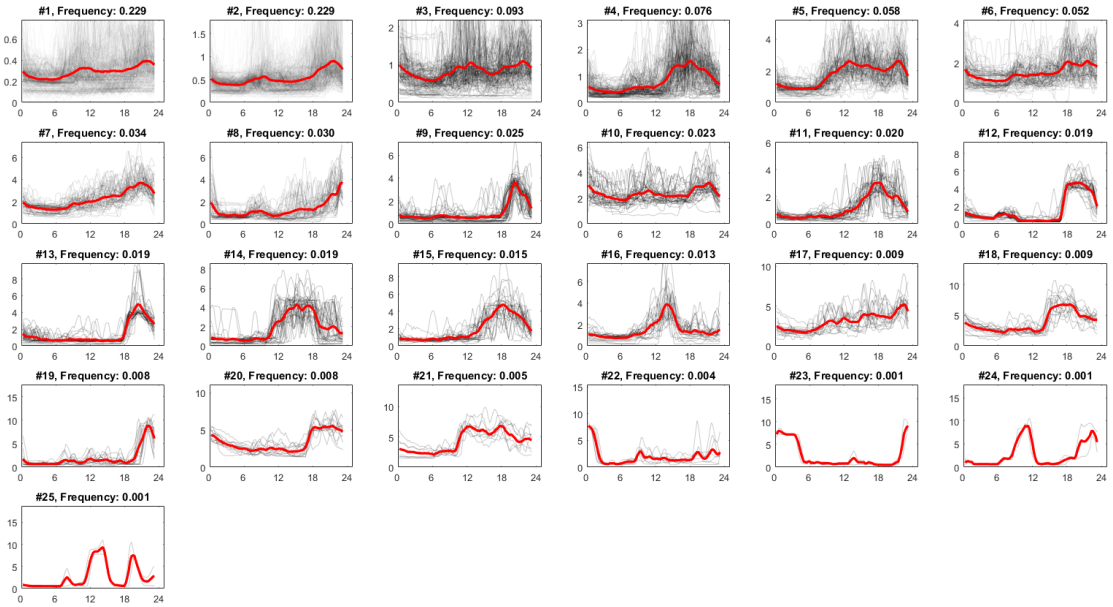


(c)

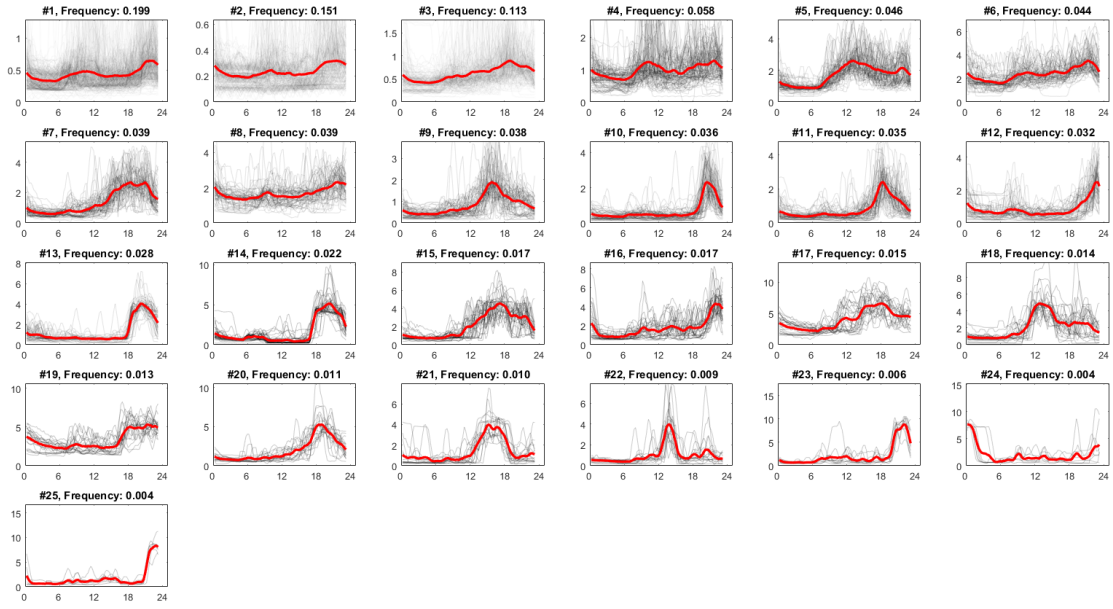
Figure 4-5. Cluster for dataset 2: (a) K-means, K=10, (b) HC, K=30, (c) SOM, K=30. The vertical axis is ‘Power (kW)’ and the horizontal axis is ‘Time (hr)’.



(a)



(b)



(c)

Figure 4-6. Cluster for dataset 3: (a) K-means, K=10, (b) HC, K=30, (c) SOM, K=30. The vertical axis is ‘Power (kW)’ and the horizontal axis is ‘Time (hr)’.

Chapter 5: Two-stage clustering on household electricity load shapes

This chapter presents a two-stage clustering for household electricity load shapes. In Chapter 4, it was shown that CVIs can result in an unreasonably low number of clusters, with clusters that conceal the actual energy load shapes in terms of temporal shape and power magnitude. On the other hand, selecting a high number of clusters could improve the representation of energy load shapes. However, it could limit the interpretability purpose and result in highly correlated clusters that are redundant. To this end, the purpose of this chapter is to present a two-stage clustering approach that initially overpopulates clusters to improve the accuracy in load shape representation and further merge the closely related ones to improve the interpretability.

5.1. Two-stage clustering

The general framework is shown in Figure *5-1*. The objective is to efficiently reduce the cluster library size after generating a high number without losing the essential information in load shape representation. In the first stage, a large number of clusters is generated, which are then transformed with an efficient time-series averaging technique to represent cluster centroids. In the second stage, pairwise distances of clusters are calculated with a distance measure that accounts for shape alignment between cluster centroids to merge the closely related ones.

In what follows, the description for each part is presented.

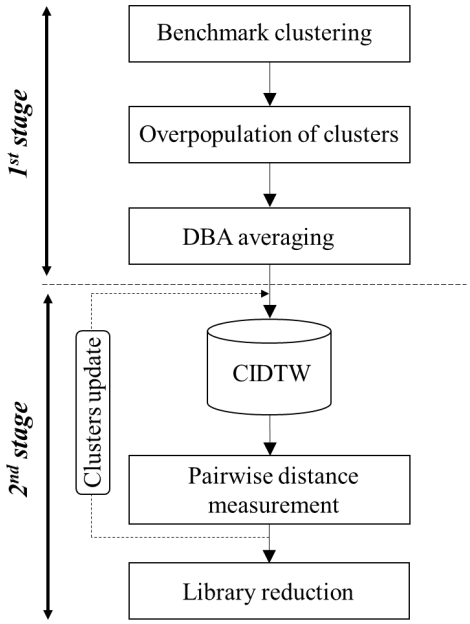


Figure 5-1. .Cluster library reduction framework

5.1.1. First stage: Initial cluster representation

5.1.1.a. Overpopulation of clusters

In the first stage, clusters are overpopulated to ensure a reasonable representation of load shapes. An initial value of K' for the number of clusters is considered ($K' > K$). The selection of K' can be done as a measurement of the sum of squared error, such that increasing K' does not cause considerable change in the SSE. To this end, the elbow curves for WCSS (described in Section 4.2) is used for the selection of K'. Using an arbitrary clustering technique, K' groups are populated.

5.1.1.b. Cluster representation

In order to merge the closely related clusters, it is essential to have a reasonable representation for each cluster. Since each cluster may include thousands of observations, it is not feasible to involve each observation in the merging process. Therefore, a proper representation for each cluster, that resembles the content of each groups is required. The most intuitive way for cluster representation is the Euclidean averaging (simple averaging of each sample across all observations). However, Euclidean averaging may result in a centroid which is dissimilar to any of its associated time-series in the associated cluster [47]. To this end, we employed the DTW Barycenter Averaging (DBA) technique [48]. DBA is a time-series averaging technique, in which the resultant centroid of each cluster is a reorientation of each group. In contrast to conventional time-series averaging which may extract a centroid that differs in shape from its original time-series, DBA uses an expectation-maximization approach by refining the medoid of each group through finding the best set of alignments within each group through iterations. Technically, in each iteration [48]:

(1) DTW distance between each profile and the temporary average centroid (which is updated in each iteration) is measured. This is performed to find the association of each sample of the centroids with the samples of the set of profiles.

(2) Updating each sample of the centroid as the barycenter of samples associated with it from the previous step.

The barycenter function is defined as:

$$barycenter(X_1, X_2, \dots, X_m) = \frac{X_1 + X_2 + \cdots + X_m}{m} \quad (12)$$

Using the average centroid from the previous iteration (C_i), the t-th sample of the average centroid in the current iteration (C_i') is defined as:

$$C_i'(t) = barycenter\left(assoc\left(C_i(t)\right)\right) \quad (13)$$

Here, the $assoc$ function associates each sample of the C_i to the samples (one or more) of the profiles during the DTW calculation.

Considering K' initial clusters, DBA was applied to the content of each to obtain the centroids C_i. To demonstrate the impact of conventional averaging and DBA, we presented several clusters in the dataset in Figure *5-2*. As can be seen, DBA could enhance the representation of centroids in each cluster.

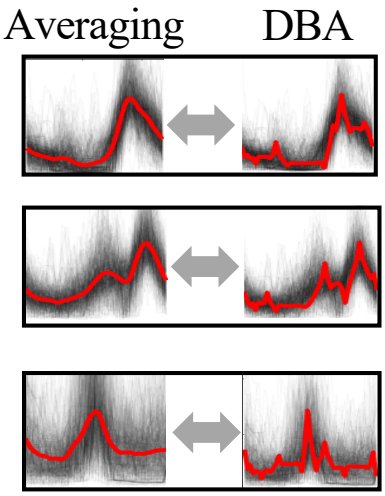


Figure 5-2. Examples of clusters with DBA averaging compared to conventional averaging.

5.1.2. Second stage: Cluster merging

In this stage, the clusters that are highly similar in shape and magnitude are merged to reduce the final library size. Given the initial cluster size of K' in the first stage, the objective is to reduce the library size to K. We used an iterative merging process to transform the dataset from K' to K clusters. In each iteration, the matrix of similarity measure between cluster i and j is constructed and the closest ones are merged ($K' \rightarrow K' - 1$). The process is continued till $K \rightarrow K'$.

Due to the nature of the electricity profile dataset, it is important to employ a robust measure for merging the time-series that accounts for the inherent relatively small time shift in similar load shapes. Figure *5-3* shows an example of two household load shapes that have similar energy behavior patterns (double demand peaks in morning and evening) but are relatively different in the time domain (~1-hour difference in peak timing). Typical similarity metrics may fail to capture the similarity of such cases. To this end, we employed the Complexity-Invariant Dynamic Time

Warping (CI-DTW) as the distance measure [49]. CI-DTW is a DTW-based distance measure that is invariant to the complexity of time-series (e.g., number of peaks and valleys). Therefore, it avoids the shortcoming of matching pairs of simple objects that are subjectively apart from those with more complex patterns with similar shapes [49].

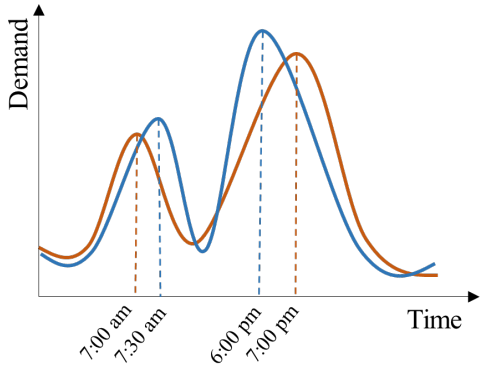


Figure 5-3. Household load shapes with similar behaviors and temporal shifts.

5.1.2.a. CI-DTW

CI-DTW is a variation of the DTW distance measure. DTW [50] recursively finds the optimal alignment of the two time-series ($P = \{p_1, p_2, \dots, p_T\}$ and $Q = \{q_1, q_2, \dots, q_T\}$) by calculating the cost defined by:

$$D(P_i, Q_j) = \delta(p_i, q_i) + \min \begin{cases} D(P_{i-1}, Q_{j-1}) \\ D(P_i, Q_{j-1}) \\ D(P_{i-1}, Q_j) \end{cases} \tag{14}$$

in which $\delta(p_i, q_i)$ is the distance between samples. The above equation can be efficiently found through dynamic programming. Figure *5-4* shows an example of how the above equation is updated in the matrix that maps the value from two time-series. The value on the upper-right side of the matrix denotes the DTW distance.

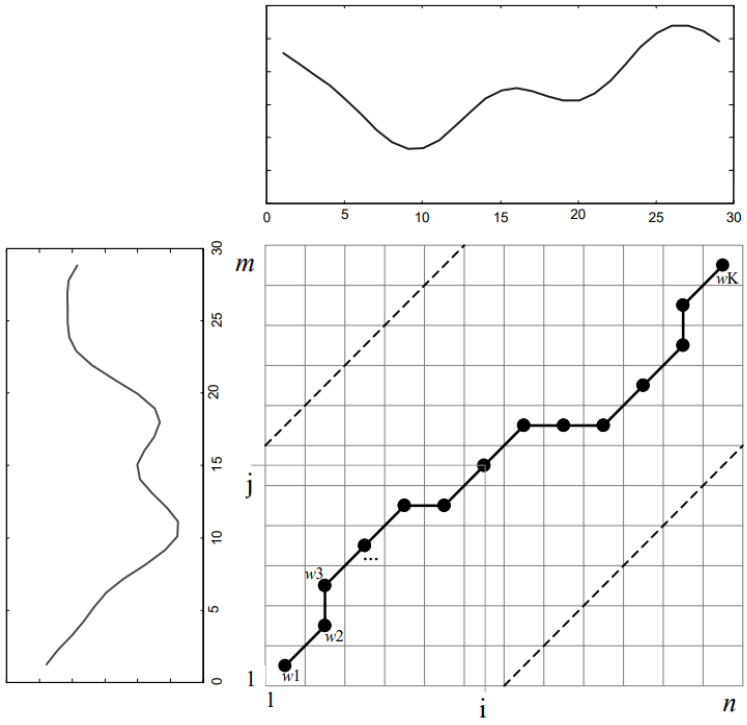


Figure 5-4. Warping path structure for measuring DTW. Taken from reference [51].

Upon measuring the DTW distance, CI-DTW is defined as:

$$CIDTW = DTW(P,Q) * CF(P,Q) \tag{15}$$

in which $CF(P,Q)$ is a correction factor as follows:

$$CF(P,Q) = \frac{\max\big(CE(P), CE(Q)\big)}{\min\big(CE(P), CE(Q)\big)} \tag{16}$$

$CE(a)$ is a complexity estimate as:

$$CE(A) = \sqrt{\sum_{i=1}^{t-1} (a_i - a_{i+1})^2} \tag{17}$$

5.1.2.b. Iterative merging

Using the distance measurement described in Section 5.1.2.a between cluster centroids, the closest pairs of clusters (i, j) are merged in each iteration. A control parameter (τ) is considered as the maximum cluster density upon merging i, j such that:

$$\|C_i\| + \|C_j\| \leq \tau * n \tag{18}$$

in which $\|C_i\|$ and $\left\|C_j\right\|$ are the number of profiles associated with clusters i and j, respectively, and n is the total number of profiles in the dataset. Therefore, τ controls the cluster size after merging to avoid the formation of clusters that are highly dense. In the presented results, $\tau = 0.2$ was considered.

Figure *5-5* shows the pseudocode for cluster merging in the second stage.

Algorithm 1. Merging cluster centroids

Input: Overpopulated clustering results, cluster centroids with DBA, initial cluster number (K')

Set the target cluster number K
While K'>K:
    Find the closest cluster centroids C_i and C_j based on CI-DTW metric.
    While $\|C_i\| + \left\|C_j\right\|$>$\tau * n$:
        Find the next set of closest C_i and C_j.
    Set $C_i = (n_i C_i + n_j C_j)/(n_i + n_j)$
    Delete C_j
    Update cluster index from cluster $j + 1$ to the last one.
    $K' = K'$-1

Figure 5-5. Pseudocode for cluster merging.

5.2. Results and discussion

5.2.1. Visualization and empirical investigation

Using the approach described in Section 5.1, the two-stage clustering was applied to the datasets. We used SOM and K-means for presenting the results. Due to the lower performance of fuzzy c-means and the tendency of hierarchical clustering to generate outliers, they were not considered. To estimate the initial number of clusters (K'), the elbow curves for WCSSE in Figure *4-1*, Figure *4-2*, and Figure *4-3* were used. A set of $K' = \{50, 70, 90\}$, spanning the range in which the error declines was low, was considered for the overpopulation of clusters. For the second stage, the final library size of $K = \{10, 20, 30, 40\}$ was used to study the impact of cluster merging at different levels.

Figure *5-6* shows the pairs of merged clusters at different iterations (K'=90; K=40; number of iterations=50). In this figure, SOM was applied for creating the initial cluster library. The value above each subplot is the iteration number. As can be seen, the results mainly show that the selected clusters are subjectively close both in temporal shape and peak magnitudes.

Upon merging, the final clusters for this example are presented in Figure 5-7. Empirical observations show that clusters are well-separated and distinct in their temporal shapes and power magnitude. Furthermore, they accentuate the useful features in load shapes such as peak magnitude, peak timing, peak duration, and energy volume, which could be of interest to utilities for designing energy programs.

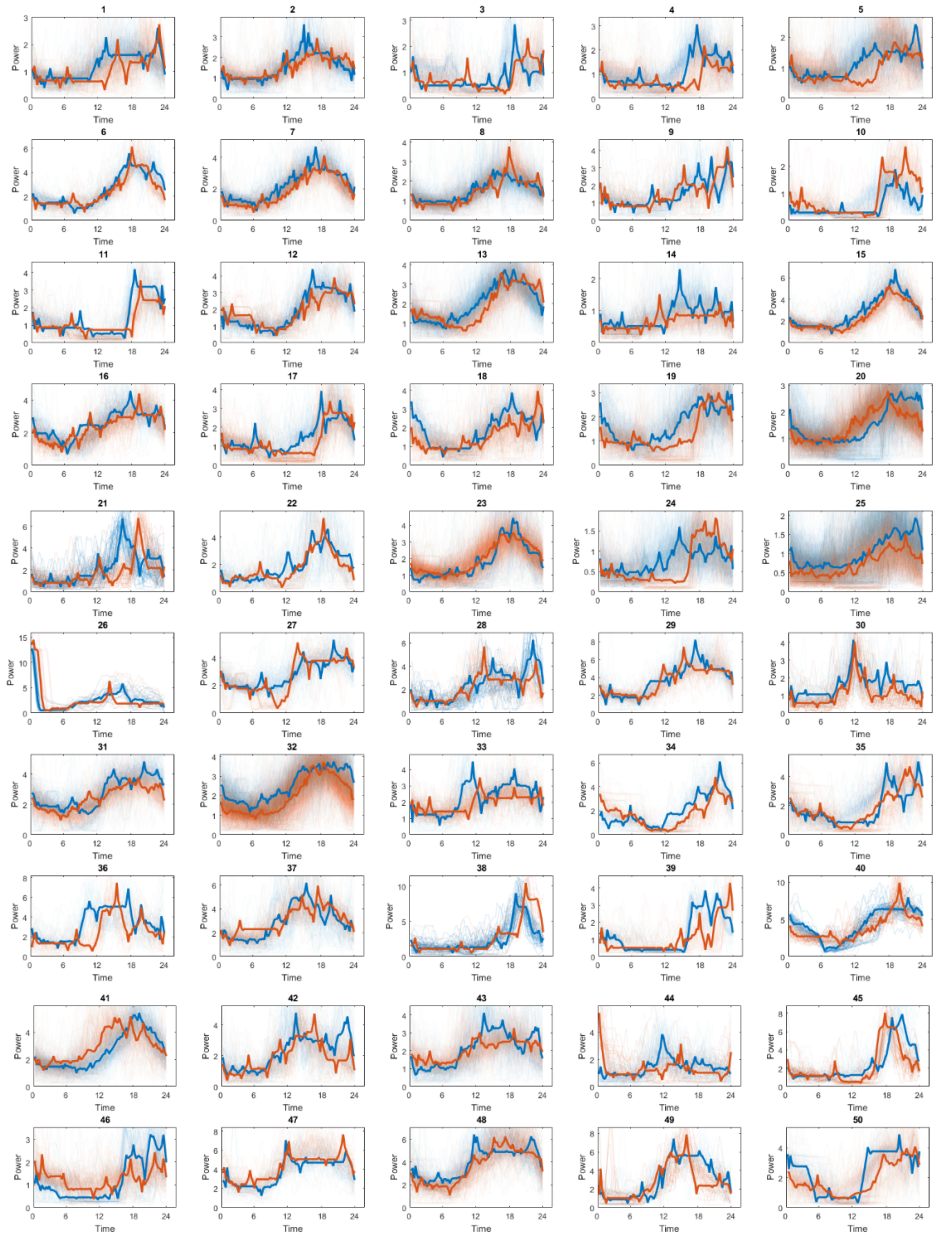


Figure 5-6. Pairs of merged clusters at different iterations (Initial library size =90 clusters; Final library size (stopping criterion) =40 clusters; Number of iterations = 50). The number above each subplot is the iteration.

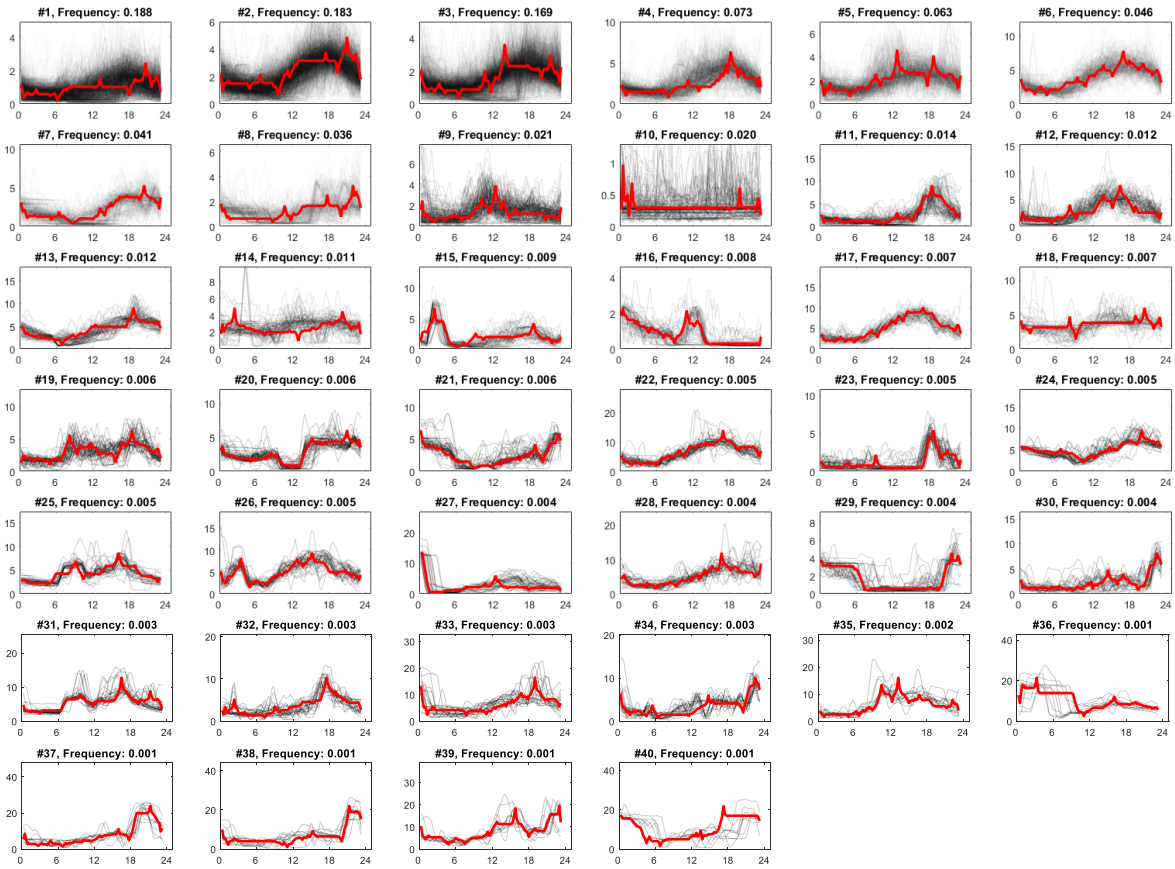


Figure 5-7. Two-stage clustering ($K' = 90, K = 40$). For all subplots, the horizontal axis is the hour (time of the day), and the vertical axis is power (kW).

5.2.2. Quantified investigation

- **Metrics:** We used WCSSE and the weighted average correlation coefficient as the two quantified metrics for comparison. Both of these values reflect the compactness of clusters. Therefore, they resemble to what extent the cluster centroids are representing their associated profiles. WCSSE has been defined in Section 4.2, and a lower value of WCSSE indicates higher compactness. Weighted average correlation (WAC) first measures the correlation coefficient of each cluster centroid with its associated profiles and uses the average as the correlation indicator of each cluster. Thereafter, it uses the density of each cluster to have the weighted average as the single correlation value. More specifically:

$$WAC = \frac{\sum_{i=1}^{K} \|C_i\| * corr_i}{N}, WAC \in [0,1] \quad (19)$$

in which WAC is the weighted average correlation and $corr_i$ is the average correlation coefficient for cluster i as follows:

$$corr_i = \frac{\sum_{x \epsilon C_i} corr(x, \mu_i)}{\|C_i\|} \tag{20}$$

A higher value of WAC indicates higher compactness.

Figure *5-8* presents the weighted average correlation coefficient for different scenarios. Each row represents one dataset, and each column is one clustering method (SOM on left and K-means on the right). The two-stage bars show the final results after overpopulating the clusters (with K' values shown in the subplots) and then merging them up to K clusters (shown on the horizontal axis). The benchmark bar represents the conventional clustering by directly selecting the K as shown in the horizontal axis. As the results show, using the two-stage approach improves the correlation in most cases. Specifically, it improves the average correlation by 8.2%, 8.9%, and 2.6% for dataset 1, dataset 2, and dataset 3, respectively.

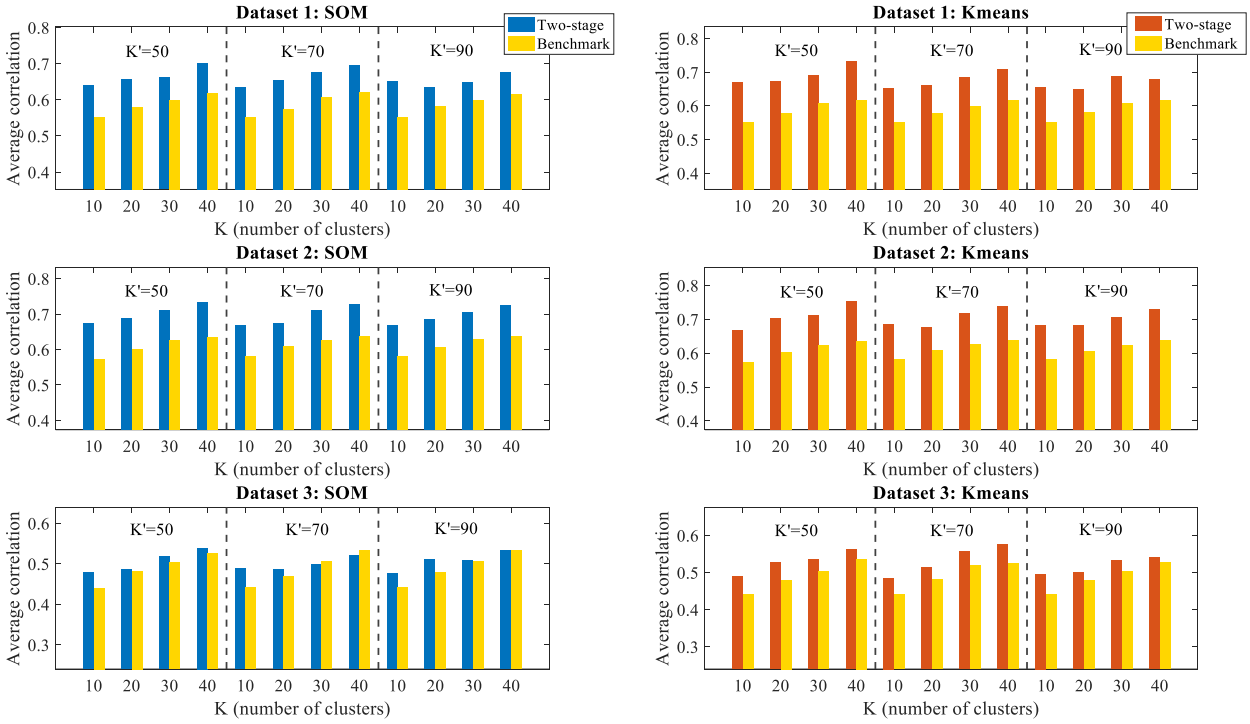


Figure 5-8. Comparison of the weighted average correlation coefficient between two-stage and benchmark clustering. The left subplots use SOM and the right subplots use K-means. The higher is better.

Similar to what was presented in Figure *5-8*, the results for WCSSE metric are illustrated in Figure *5-9*. As shown, the two-stage approach results have lower error value on average, therefore,

indicating higher compactness of clusters. Specifically, it reduces the WCSSE average by 9.3%, 9.5%, for dataset 1 and dataset 2, respectively. However, for dataset 3, an average of 3.4% increase was observed. A possible interpretation for the error increase for this case is that the size of dataset 3 was considerably lower compared to other datasets (see Table *3-1*). Therefore, the initial K' values used in Figure *5-9* may not be appropriate for this dataset. To this end, the solutions by recent efforts [52] for selecting the appropriate cluster number might address this issue.

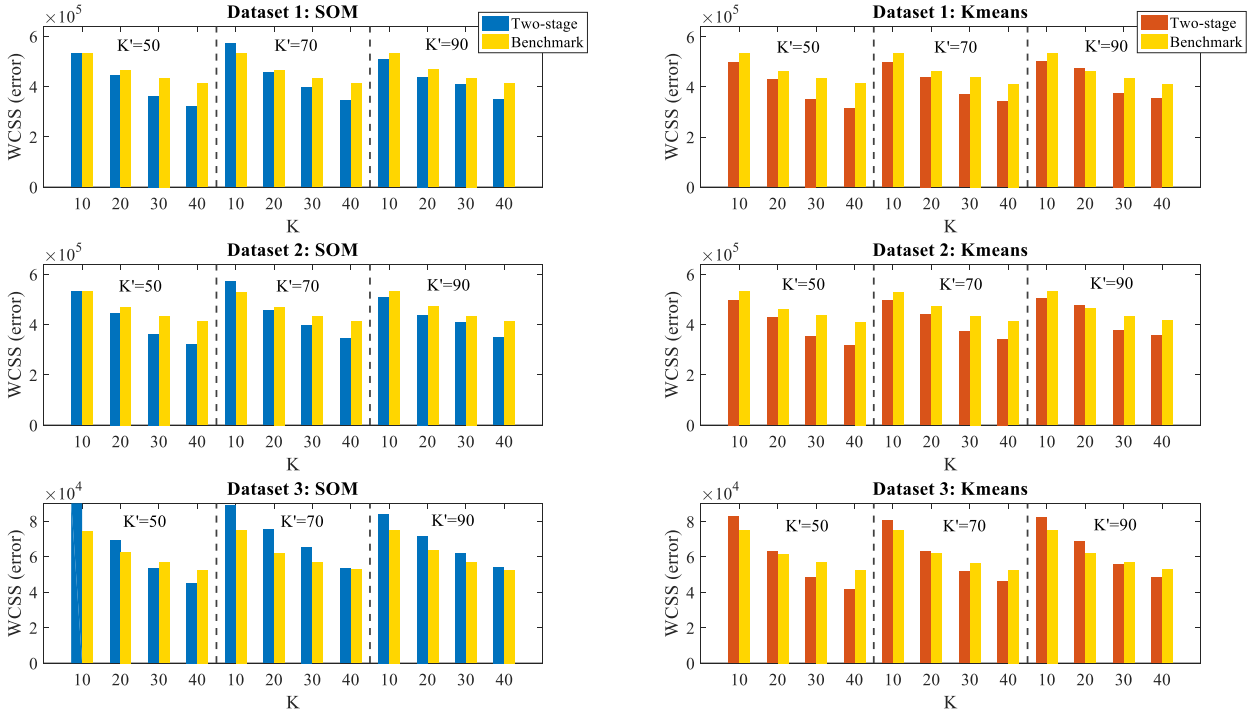


Figure 5-9. Comparison of the WSSE between two-stage and benchmark clustering. The left subplots use SOM and the right subplots use K-means. The lower is better.

5.3. Impact

5.3.1. Applications to Energy program

Based on the empirical and quantified evaluation in Section 5.2, the proposed two-stage clustering could help to accentuate the useful features in load shapes such as peak magnitude, peak timing, peak duration, and energy volume, which are important to electricity providers. To provide some context on the implication of findings, we interpreted the results in Figure *5-7* from the application standpoint.

Demand Response (DR) events are typically scheduled for the evening time frame, in which the net demand of the network is excessively high. Based on Figure *5-7*, clusters 4, 5, 6, 7, 11, 13, 17, 19, 22, 33 have sharp peaks during the 17:00-19:00 time frame, which make them potential candidates for DR events during those times. The implementation of DR could be made by partial load shedding or load shifting. Clusters 2, 8, 21, 30, 34, and 37 have also a sharp peak at a later time frame, which make them potential candidates for DR for time frames after 20:00. Especially, clusters 30, 32, 33, 34, and 37 have considerably high usage (peak demand of more than 10kW), whose peak consumption could be tied with heavy usage appliances such as pool pump, multiple AC units, and simultaneous usage of wet appliance or electric vehicle (EV) charging.

In another category for distributed energy management applications, the increased integration of renewable resources such as solar generation is of utmost importance to move towards energy decarbonization. To this end, clusters 9, 16, and 35, whose demand patterns in the noon time frame coincide with the solar generation, could benefit more from photovoltaic (PV) integration. Additionally, clusters 20 and 29 have almost zero demand at noon, implying they could save a considerable amount of energy in batteries if they have PV-battery systems. Alternatively, with the rise of new technologies such as peer-to-peer energy trading between prosumers and consumers, these clusters, if equipped with PV, could also offer their high amount of on-site generation to their consumer neighbors.

Finally, clusters 14, 15, 26, and 27 have a sharp peak after midnight, probably due to EV charging and wet appliances (e.g., dishwasher) operation. Since the time frame of heavy usage is during the off-peak demand time, their energy behavior is already compatible with DR plans, and no types of incentives would be needed (assuming the typical evening/night peak demand for the network).

5.3.2. New customer classification

The high variation of energy usage causes an increased number of distinct load shapes and clusters. However, given the existence of adequate data that captures different possibilities in energy load shapes and magnitude, a pre-processed cluster library can be created, which can assist for classifying new customers. Accordingly, when a new customer is added to the community, their load shapes could be either classified into an existing cluster or generating a new one. For example, a simple KNN algorithm could identify the most similar clusters to the load shapes of a new customer. Thereafter, metrics such as the correlation between the cluster centroid in the pre-processed library and the new load shape could decide whether the new observation is associated

with the existing dataset. For example, if correlation falls below a certain value, it could suggest the creation of a new cluster or an outlier observation.

Chapter 6: Conclusion and future directions

With the increased adoption of smart metering infrastructures, a large amount of time-series energy data is generated, which provides opportunities for customers' energy behavior analytics. Clustering is deemed as a proper approach for segregating the large pool of energy load shapes into a limited number of representative patterns, which could further be used by utilities for resource allocation and program design. However, many clustering techniques may result in an oversimplified representation of load shapes or obtaining clusters whose centroids deviate considerably from their associated load shapes. In this thesis, we introduced a two-stage clustering to preserve the temporal patterns and magnitude of load shapes for consumer segmentation. The contribution of this work is two-fold: (1) We presented a comparative assessment of different conventional clustering techniques and CVIs. We further showed that relying on common CVIs can lead to false predictions in the cluster number and such generic metrics are not suited to this specific problem, (2) We introduced a two-stage clustering approach with improved representation in temporal shape and energy volume. Empirical investigation and quantified assessment compared to the benchmark solutions were presented to demonstrate the applicability of the method.

Impact. This work introduced a clustering approach, and it focused on the application of building energy profile segmentation. Although the findings were presented for a specific task in the building science domain, it is worthwhile to investigate this approach in other domains. Specifically, the primary motivation was to distinguish between the locations of peaks and valleys in time-series in addition to highlighting the peak magnitudes. Therefore, exploring the findings on datasets with similar natures in other domains, such as finance or healthcare, could be interesting.

Future directions. Considering the high-granularity of load shape segmentation proposed in this work, several interesting directions for future research are elaborated as follows:
(1) With the increased adoption of datasets, pre-processed cluster libraries can be created. Therefore, by adding new customers whose energy behavior has not been previously seen in the community, one could classify the new customer into one of the representative clusters.

Furthermore, by measuring the accuracy of the new customer's daily profile with respect to the nearest cluster in the pre-processed library, one could decide if the new types of energy behavior are matched with those already seen in the library, or the library should be updated with new clusters (only if the new behavior is observed regularly and has a minimum density in the database).

(2) One interesting direction of load shape segmentation is to infer the possible time-of-use for different categories of appliances. Accordingly, the sharp peaks observed in cluster centroids could be associated with the usage of high power-draw devices in buildings, such as pool pump, dryer, or EV charging. Therefore, through using appliance-level data tied with the smart meter data, we could investigate the possible drivers of consumption for daily load shapes (e.g., [53]). For example, one can use time-series filtering to deconvolute changes in the total consumption signal into signals reflecting the activation time of various appliances. To accomplish this, one can extract baseline measures of confounding variables from periods of low consumption activity (e.g., refrigerator, average cooling/heating system) and modify the total consumption data by filtering out contributions from the confounding variables. Furthermore, one can use temporal sequence abstraction to identify events (times associated with steps or changes in consumption activity) into temporal ranges corresponding to different appliances or combinations of appliances. Accordingly, future directions could comprise of: (I) developing supervised models by integrating appliance-level data for a set of buildings to make predictions for a set of unknown buildings, (II) investigating correlation analysis between sharp peaks of different clusters and different categories of appliances with high-power usage.

(3) In this thesis, we used a sample of data from the Pecan Street Dataport [42], which is the largest publicly available residential energy database. In this dataset, different buildings contained a variety of different sets of appliances (e.g., AC, wet appliances, EV, furnace, and pool pump). Furthermore, the building types were categorized as a single-family house, apartment, or townhouse. Given this variation, the dataset can be regarded as heterogeneous in terms of energy consumption patterns. However, further investigation is needed for future research on a larger spatial and temporal scale. Specifically, several factors to explore, which add more variations to datasets include the impact of occupancy (number of occupants), working lifestyle (full-time, part-time, retired), age of occupants, and the presence of different appliances.

(4) One important aspect of big data analytics is the computational efficiency and runtime. For the introduced two-stage technique, the first stage uses benchmark algorithms, many of which are computationally efficient. For example, the complexity of the K-means is $O(tknd)$, where t is the number of iteration, k is the number of clusters, n is the number of observations, and d is the dimension. The second stage, though, employs DTW, which has a $O(n^2)$ time and space complexity, and DBA, which has an overall complexity of $O(tnd^2)$. Accordingly, the calculations of the second stage could be a bottleneck for datasets containing hundreds of thousands of observations. Therefore, future research could aim at integrating more computationally efficient alternates towards big data analytics. To this end, examples of faster variations for DTW calculations such as [54, 55] can be integrated and tested in the future works.

References

[1] U. S. Energy Information Administration. Electricity overview. (2017). Available: https://www.eia.gov/electricity/data.php#summary. Last retreived on June 17, 2020.

[2] F. McLoughlin, A. Duffy, and M. Conlon, "A clustering approach to domestic electricity load profile characterisation using smart metering data," *Applied energy,* vol. 141, pp. 190-199, 2015.

[3] G. J. Tsekouras, N. D. Hatziargyriou, and E. N. Dialynas, "Two-stage pattern recognition of load curves for classification of electricity customers," *IEEE Transactions on Power Systems,* vol. 22, no. 3, pp. 1120-1128, 2007.

[4] G. Chicco, "Overview and performance assessment of the clustering methods for electrical load pattern grouping," *Energy,* vol. 42, no. 1, pp. 68-80, 2012.

[5] G. Tsekouras, P. Kotoulas, C. Tsirekis, E. Dialynas, and N. Hatziargyriou, "A pattern recognition methodology for evaluation of load profiles and typical days of large electricity customers," *Electric Power Systems Research,* vol. 78, no. 9, pp. 1494-1510, 2008.

[6] V. Figueiredo, F. Rodrigues, Z. Vale, and J. B. Gouveia, "An electric energy consumer characterization framework based on data mining techniques," *IEEE Transactions on power systems,* vol. 20, no. 2, pp. 596-602, 2005.

[7] D. Hsu, "Comparison of integrated clustering methods for accurate and stable prediction of building energy consumption data," *Applied Energy,* vol. 160, pp. 153-163, 2015.

[8] G. Chicco, R. Napoli, and F. Piglione, "Comparisons among clustering techniques for electricity customer classification," *IEEE Transactions on Power Systems,* vol. 21, no. 2, pp. 933-940, 2006.

[9] T. Räsänen, D. Voukantsis, H. Niska, K. Karatzas, and M. Kolehmainen, "Data-based method for creating electricity use load profiles using large amount of customer-specific hourly measured electricity use data," *Applied Energy,* vol. 87, no. 11, pp. 3538-3545, 2010.

[10] G. Chicco, R. Napoli, F. Piglione, P. Postolache, M. Scutariu, and C. Toader, "Load pattern-based classification of electricity customers," *IEEE Transactions on Power Systems,* vol. 19, no. 2, pp. 1232-1239, 2004.

[11] I. P. Panapakidis, T. A. Papadopoulos, G. C. Christoforidis, and G. K. Papagiannis, "Pattern recognition algorithms for electricity load curve analysis of buildings," *Energy and Buildings,* vol. 73, pp. 137-145, 2014.

[12] J. Kwac, J. Flora, and R. Rajagopal, "Household energy consumption segmentation using hourly data," *IEEE Transactions on Smart Grid,* vol. 5, no. 1, pp. 420-430, 2014.

[13] B. Yildiz, J. Bilbao, J. Dore, and A. Sproul, "Recent advances in the analysis of residential electricity consumption and applications of smart meter data," *Applied Energy,* vol. 208, pp. 402-427, 2017.

[14] H. Hino, H. Shen, N. Murata, S. Wakao, and Y. Hayashi, "A versatile clustering method for electricity consumption pattern analysis in households," *IEEE Transactions on Smart Grid,* vol. 4, no. 2, pp. 1048-1057, 2013.

[15] J. D. Rhodes, W. J. Cole, C. R. Upshaw, T. F. Edgar, and M. E. Webber, "Clustering analysis of residential electricity demand profiles," *Applied Energy,* vol. 135, pp. 461-471, 2014.

[16] Y. Wang, Q. Chen, C. Kang, and Q. Xia, "Clustering of electricity consumption behavior dynamics toward big data applications," *IEEE transactions on smart grid,* vol. 7, no. 5, pp. 2437-2447, 2016.
[17] S. Haben, C. Singleton, and P. Grindrod, "Analysis and clustering of residential customers energy behavioral demand using smart meter data," *IEEE transactions on smart grid,* vol. 7, no. 1, pp. 136-144, 2016.
[18] O. Motlagh, A. Berry, and L. O'Neil, "Clustering of residential electricity customers using load time series," *Applied energy,* vol. 237, pp. 11-24, 2019.
[19] M. Afzalan and F. Jazizadeh, "Residential loads flexibility potential for demand response using energy consumption patterns and user segments," *Applied Energy,* vol. 254, p. 113693, 2019.
[20] S. Xu, E. Barbour, and M. C. González, "Household segmentation by load shape and daily consumption," in *Proc. ACM SigKDD 2017 Conf. Halifax, Nov. Scotia, Canada, August 2017*, 2017.
[21] A. Kavousian, R. Rajagopal, and M. Fischer, "Determinants of residential electricity consumption: Using smart meter data to examine the effect of climate, building characteristics, appliance stock, and occupants' behavior," *Energy,* vol. 55, pp. 184-194, 2013.
[22] K. Zhou and S. Yang, "Understanding household energy consumption behavior: The contribution of energy big data analytics," *Renewable and Sustainable Energy Reviews,* vol. 56, pp. 810-819, 2016.
[23] S. Iyengar, S. Lee, D. Irwin, and P. Shenoy, "Analyzing energy usage on a city-scale using utility smart meters," in *Proceedings of the 3rd ACM International Conference on Systems for Energy-Efficient Built Environments*, 2016, pp. 51-60.
[24] T. Zhang, G. Zhang, J. Lu, X. Feng, and W. Yang, "A new index and classification approach for load pattern analysis of large electricity customers," *IEEE Transactions on Power Systems,* vol. 27, no. 1, pp. 153-160, 2012.
[25] S. V. Verdú, M. O. Garcia, C. Senabre, A. G. Marin, and F. G. Franco, "Classification, filtering, and identification of electrical customer load patterns through the use of self-organizing maps," *IEEE Transactions on Power Systems,* vol. 21, no. 4, pp. 1672-1682, 2006.
[26] G. Coke and M. Tsao, "Random effects mixture models for clustering electrical load series," *Journal of time series analysis,* vol. 31, no. 6, pp. 451-464, 2010.
[27] F. L. Quilumba, W.-J. Lee, H. Huang, D. Y. Wang, and R. L. Szabados, "Using smart meter data to improve the accuracy of intraday load forecasting considering customer behavior similarities," *IEEE Transactions on Smart Grid,* vol. 6, no. 2, pp. 911-918, 2015.
[28] G. Chicco, R. Napoli, P. Postolache, M. Scutariu, and C. Toader, "Customer characterization options for improving the tariff offer," *IEEE Transactions on Power Systems,* vol. 18, no. 1, pp. 381-387, 2003.
[29] A. D. Fontanini and J. Abreu, "A Data-Driven BIRCH Clustering Method for Extracting Typical Load Profiles for Big Data," in *2018 IEEE Power & Energy Society General Meeting (PESGM)*, 2018, pp. 1-5: IEEE.
[30] J. Kwac, J. Flora, and R. Rajagopal, "Lifestyle segmentation based on energy consumption data," *IEEE Transactions on Smart Grid,* vol. 9, no. 4, pp. 2409-2418, 2018.

[31] T. Teeraratkul, D. O'Neill, and S. Lall, "Shape-based approach to household electric load curve clustering and prediction," *IEEE Transactions on Smart Grid,* vol. 9, no. 5, pp. 5196-5206, 2018.
[32] D. J. Berndt and J. Clifford, "Using dynamic time warping to find patterns in time series," in *KDD workshop*, 1994, vol. 10, no. 16, pp. 359-370: Seattle, WA.
[33] E. Barbour and M. González, "Enhancing household-level load forecasts using daily load profile clustering," in *Proceedings of the 5th Conference on Systems for Built Environments*, 2018, pp. 107-115: ACM.
[34] M. Hu and F. Xiao, "Investigation of the demand response potentials of residential air conditioners using grey-box room thermal model," *Energy Procedia,* vol. 105, pp. 2759-2765, 2017.
[35] S. Lin, F. Li, E. Tian, Y. Fu, and D. Li, "Clustering load profiles for demand response applications," *IEEE Transactions on Smart Grid,* 2017.
[36] A. Rajabi, M. Eskandari, M. J. Ghadi, L. Li, J. Zhang, and P. Siano, "A comparative study of clustering techniques for electrical load pattern segmentation," *Renewable and Sustainable Energy Reviews,* vol. 120, p. 109628, 2020.
[37] S. Haben, C. Singleton, and P. Grindrod, "Analysis and clustering of residential customers energy behavioral demand using smart meter data," *IEEE transactions on smart grid,* vol. 7, no. 1, pp. 136-144, 2015.
[38] M. Afzalan and F. Jazizadeh, "Semantic search in household energy consumption segmentation through descriptive characterization," in *Proceedings of the 6th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*, 2019, pp. 263-266.
[39] A. Malik, N. Haghdadi, I. MacGill, and J. Ravishankar, "Appliance level data analysis of summer demand reduction potential from residential air conditioner control," *Applied Energy,* vol. 235, pp. 776-785, 2019.
[40] S. Lin, F. Li, E. Tian, Y. Fu, and D. Li, "Clustering load profiles for demand response applications," *IEEE Transactions on Smart Grid,* vol. 10, no. 2, pp. 1599-1607, 2017.
[41] J. Kwac, J. Flora, and R. Rajagopal, "Lifestyle segmentation based on energy consumption data," *IEEE Transactions on Smart Grid,* vol. 9, no. 4, pp. 2409-2418, 2016.
[42] Source: Pecan Street Dataport. 2017. [Online]. Available: https://dataport.pecanstreet.org/. Last retrieved on June 17, 2020
[43] Y. Liu, Z. Li, H. Xiong, X. Gao, and J. Wu, "Understanding of internal clustering validation measures," in *2010 IEEE International Conference on Data Mining*, 2010, pp. 911-916: IEEE.
[44] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE transactions on pattern analysis and machine intelligence,* no. 2, pp. 224-227, 1979.
[45] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of computational and applied mathematics,* vol. 20, pp. 53-65, 1987.
[46] T. Caliński and J. Harabasz, "A dendrite method for cluster analysis," *Communications in Statistics-theory and Methods,* vol. 3, no. 1, pp. 1-27, 1974.
[47] F. Petitjean, G. Forestier, G. I. Webb, A. E. Nicholson, Y. Chen, and E. Keogh, "Dynamic time warping averaging of time series allows faster and more accurate classification," in *2014 IEEE international conference on data mining*, 2014, pp. 470-479: IEEE.

[48] F. Petitjean, A. Ketterlin, and P. Gançarski, "A global averaging method for dynamic time warping, with applications to clustering," *Pattern Recognition,* vol. 44, no. 3, pp. 678-693, 2011.
[49] G. E. Batista, X. Wang, and E. J. Keogh, "A complexity-invariant distance measure for time series," in *Proceedings of the 2011 SIAM international conference on data mining*, 2011, pp. 699-710: SIAM.
[50] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE transactions on acoustics, speech, and signal processing,* vol. 26, no. 1, pp. 43-49, 1978.
[51] E. J. Keogh and M. J. Pazzani, "Derivative dynamic time warping," in *Proceedings of the 2001 SIAM international conference on data mining*, 2001, pp. 1-11: SIAM.
[52] N. Al Khafaf, M. Jalili, and P. Sokolowski, "A Novel Clustering Index to Find Optimal Cluster Size with Application to Segmentation of Energy Consumers," *IEEE Transactions on Industrial Informatics,* 2020.
[53] M. Afzalan and F. Jazizadeh, "A Machine Learning Framework to Infer Time-of-Use of Flexible Loads: Resident Behavior Learning for Demand Response," *IEEE Access,* vol. 8, 2020.
[54] S. Salvador and P. Chan, "Toward accurate dynamic time warping in linear time and space," *Intelligent Data Analysis,* vol. 11, no. 5, pp. 561-580, 2007.
[55] D. F. Silva and G. E. Batista, "Speeding up all-pairwise dynamic time warping matrix calculation," in *Proceedings of the 2016 SIAM International Conference on Data Mining*, 2016, pp. 837-845: SIAM.