Lappeenranta University of Technology
School of Engineering Science
Computational Engineering and Technical Physics
Intelligent Computing

Master's Thesis

Kalifa Manjang

IDENTIFICATION OF CUSTOMER PROFILES FROM ELECTRICITY CONSUMPTION DATA

Examiners: Prof. Lasse Lensu
Assoc. Prof. Samuli Honkapuro

Supervisors: Adjunct Prof., Dr. Xiao-Zhi Gao
Associate Prof. Arto Kaarna
Prof. Lasse Lensu

ABSTRACT

Lappeenranta University of Technology
School of Engineering Science
Computational Engineering and Technical Physics
Intelligent Computing

Kalifa Manjang

Identification of customer profiles from electricity consumption data

Master's Thesis

2018

63 pages, 23 figures, 16 tables.

Examiners: Prof. Lasse Lensu
Assoc. Prof. Samuli Honkapuro

Keywords: K-means clustering, genetic algorithm, power user profiling, Davies-Bouldin index, Silhouette index, Calinski-Habarasz index.

The electric power suppliers are interested in identifying and categorising their consumers' profiles into different categories according to their energy consumption habits. The profiling of users can help with understanding how the users consume the energy and how the energy usage may affect the electricity distribution grid. However, the privacy of the electricity users is well protected by the current law. This study focuses on data mining methods to extract the relevant knowledge based on anonymous data obtained from smart meters. The K-means clustering algorithm was used in grouping the energy consumption data. To improve the quality of the clusters formed via the K-means clustering and to tackle the common problem of local optimum, the genetic algorithm (GA) was adopted in refining the clusters. The use of two validity indices to compare the methods showed that combining K-means and GA did indeed improve the clustering quality.

PREFACE

Bismillahi, Rahmani, Rahmeen all praise be to Allah. I would like to thank my supervisors for their undivided attention, dedication and guidance provided to me during the course of this master's thesis. This work would not have otherwise been achieved without their support. I thank my beautiful wife for her patience. To my parents, thank you for supporting my dreams of getting a higher education and instilling in me discipline, respect and hard work. Finally, I would like to thank the LUT administration for the scholarship I was offered to pursue a double degree program in this prestigious institution. I am grateful.

Lappeenranta, August 31, 2018

Kalifa Manjang

CONTENTS

LIST OF ABBREVIATIONS

BCSS	Between-Cluster Sum of Squares
CDI	Clustering Dispersion Indicator
CFSFDP	Fast Search and Find of Density Peaks
DB	Davies–Bouldin
EA	Evolutionary Algorithms
FCM	Fuzzy Clustering Means
GA	Genetic Algorithm
SAX	Symbolic Aggression Approximation
SVM	Support Vector Machine
SOM	Self Organizing Maps
WCSS	Within-cluster Sum of Squares

1 INTRODUCTION

1.1 Background

With the emergence of smart meters, more information about a user's electricity consumption can be collected easily. Prior knowledge about the group a particular user belongs to is known by the energy company to some extent. This is achieved through knowledge about the type of appliances in use or type of heating system used in the buildings. This information is stored and used in customer grouping. One shortcoming of this method is that the recorded information is seldom updated. With time, the energy consumption of the user, for example, the type of heating or electrical appliance usage might change remarkably from the known behaviours. In this regard, this single-shot method of consumer categorization is inefficient.

The traditional energy user grouping is performed using three user categories: industrial, residential and commercial users. An example of the industrial users are factories, commercial users are the shops, restaurants and supermarkets, and residential users refer to homes and apartment buildings. The consumption pattern of the energy users is much more complex than these mentioned groups [1]. The energy users should be categorized based on the pattern of electricity behaviour they exhibit.

The general application of electricity user profiling is that the knowledge of how the customers use the electricity can help the energy companies to create important policies ranging from network planning, demand response, and load forecasting [1]. A more detailed application of load profiling is described as follows [2]:

Distribution network operation

- Real-time recognition of network loadings and voltages.
- It ensures that the network is kept within its operating limits.
- It can manage post-fault supply restoration.

Short-term operation planning

- Very useful in congestion forecasting.

- Load profiles have immense application in network reconfiguration, for instance, to minimize network lost.
- Planned outages preparation.

Distribution network planning

- New base profile for probabilistic network planning.
- Ensuring the network can cost-efficiently host all foreseeable loads and generators.

Other applications of load profiling are the design of tariffs, target sales based on customers load profiles, the load profile can also be used in the adjustment of the electricity retail forecasts when new customers are contracted or old ones lost [2].

By virtue of this important demand, finding and understanding clusters using data mining techniques (scientific methods) is worthwhile [3]. Since the user identity is protected under the European Union privacy laws [4], the specific locations and identities of the participants will remain anonymous. This research applies data mining methods (clustering) to the provided data set to show users with similar energy consumption patterns and group these users together.

1.2 Objectives and delimitations

This master's thesis aims at achieving the following objectives:

- To study and use the K-means clustering algorithm for the purpose of load profiling.
- To choose and verify the appropriate number of clusters to use in the categorization of the energy users.
- To use the GA to improve the quality of the clusters formed.

This study is limited in that the results that will be obtained cannot be verified because of the anonymity of the participants.

1.3 Structure of the thesis

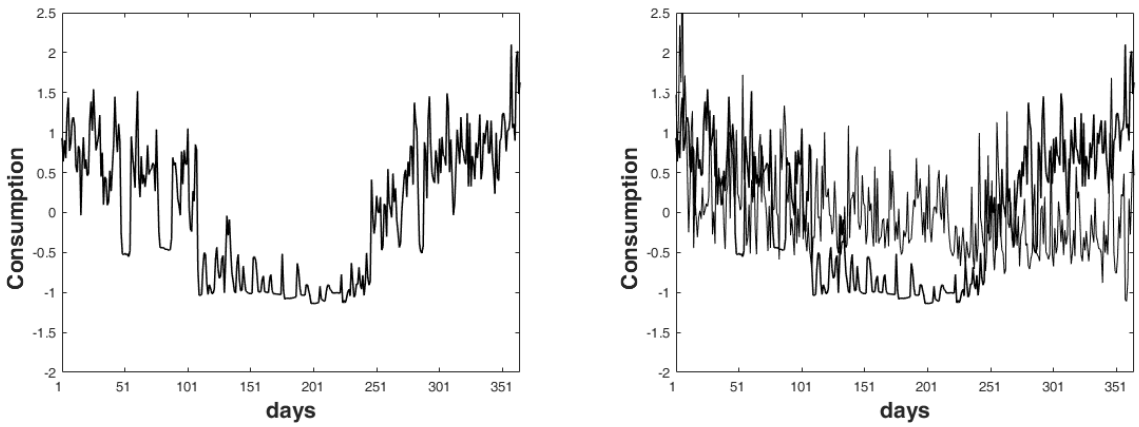
The outline of this master's thesis is as follows. Chapter 2 describes related work on power use profiling, Chapter 3 contains the proposed methods for power user profiling and the algorithms for these methods are presented. In Chapter 4, the application of these specific methods to electricity load data and the results derived from the experiment are analysed. Chapter 5 contains the discussions and the challenges faced during the study. Finally, the concluding remarks are presented in Chapter 6.

2 ELECTRICITY CONSUMER PROFILING

The tendency to precisely classify a user into the appropriate cluster is the principal challenge facing electricity user profiling. Some clustering methods (K-means clustering) require knowledge about the number of clusters prior to clustering, but this information is not always known beforehand. As a consequence of this, different clustering techniques need to be studied and the method that suits the specific purpose applied. Many studies exist already for profiling electricity users [5–7] and load profiling has proven to be one of the most suitable methods [6]. In this section, a brief summary of load profiling is given and some previous data mining approaches used in user load profiling are considered.

2.1 Power user profiling

The idea of load profiling has been around for decades already [8]. In simple terms, load profiles refer to the variation of electricity usage with time, that is how the consumers used electricity at different times during the day. In more advanced terms, a load profile is the pattern of electricity consumption displayed by a consumer or even a group of consumers over a certain period of time [9]. This period can range from half-hourly recordings, hourly recording or consumption rates across the whole year [8]. Figure 1 represents a single user load profile and multiple user load profiles respectively.



(a) (b)

Figure 1. Example load profile of residential flats: (a) A single load profile. (b) Multiple load profiles.

Furthermore, load profiling can be defined as load shapes that are determined from past

or current data and adjusted on a present or monthly basis to the actual reading provided by a meter [9].

Two main stages are involved in load profiling [10]:

1. An estimation is done on determining the groups of users over a certain period of time.
2. The information derived from the first process is then used to categorize the users.

2.2 Techniques used in power user profiling

A number of tools and techniques exist for classifying load profile data. The main design behind these techniques is to gather load profiles with similar characteristics or pattern of consumption in a cluster. The information from the clusters have a range of applications. These applications include, but not limited to billing, marketing, and utility customers [11]. These techniques are neural approaches, clustering algorithms, the statistical and fuzzy techniques. A review of these techniques is given here.

2.2.1 Neural approaches

Neural networks are very powerful tools for clustering data. This paradigm was inspired by the biological neural networks. The model first learns from the observed data and once the learning is achieved, the model is able to group new unseen data with a high level of accuracy. Neural networks have proven to be very efficient in load profiling in certain areas where other methods are not accurate [11]. For load profile division, the self-organizing map (SOM) is shown to be the most suitable [5, 12].

2.2.2 Clustering algorithms

Clustering is a type of unsupervised learning approach. It works by segmenting data points such that points that possess the same traits are grouped together distinct from other data points. The classic clustering methods are divided into Binary splitting algorithms, one step clustering, iterative joint-two algorithms and exhaustive binary search algorithms

[11]. A survey of some clustering techniques used in data mining is given in [13]. Figure 2 gives a summary of the clustering algorithms.

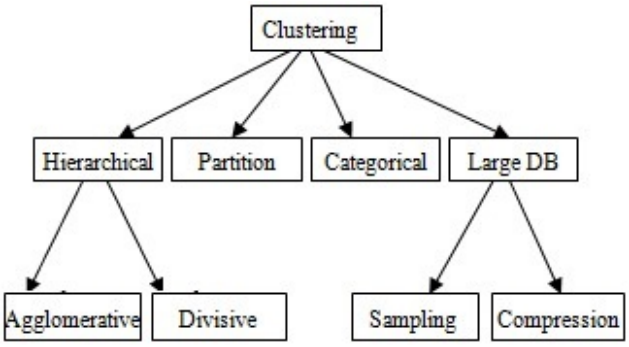


Figure 2. Clustering Techniques [14].

2.2.3 Statistical approaches

Various statistical methods have already been utilized while working with load profiles. Like the clustering techniques, the statistical methods can also serve as a useful tool in load profile data categorization. These techniques are also very valuable in addressing relationships between explaining variables (temperatures for instance) and load consumption data [11], to determine the correlation between a load profile and seasonal variations. An important question is whether statistical methods can be used in finding out what happened to the load profile in individual seasons compared to others. Statistical methods are very effective in this area.

2.2.4 Fuzzy approaches

Fuzzy clustering technique is also referred to as soft clustering. It is a type of clustering in which a data point can simultaneously belong to more than one cluster with certain membership function. This method of clustering was introduced in 1981 by Jim Bezdek [15].

2.2.5 Hybrid methods

In the review of the previous studies, it had been seen that most of the studies combine methods (two methods or more) to produce the load profiles categories [16, 17] on the same data and compare their suitability. In other general cases, one method is used to compliment another method [11].

2.3 Review of techniques

In the rest of this subsection, a review of the techniques used in related works is given.

An automatic clustering method was devised in organizing consumers into automatically created class, in which similar consumers formed the same groups. The approach was provided for characterizing customer load profiles with an aim at forming improved tariff offers for distribution companies and retailers [6].

Because of the challenge of large sample size (the enormous size of data), volatility, and uncertainty (this means the number of customers and the fact that a customer could belong to different clusters at different days), a novel method called multi-resolution clustering (MRC) to address these limitations was proposed [18]. The method was designed for smart meter data. Spectral features are extracted from the load profiles and fed as input data for clustering, computational complexity is broken down during the clustering and different load features are isolated and finally, a Gaussian mixture model is used to pre-cluster each customer over days. This is mostly to decrease the sample size. In [19], load profiling was done through K-means clustering using the SPSS statistical package. In the proposed approach, consumption data was split into 3 intervals. The motivation for splitting the 24-hour data was to capture low/high peaks of active energy consumption which is otherwise not captured in normal profiling. A comparison was made between the tariff generated from load profiles and fixed rate tariffs.

In [20] authors applied clustering to the Queensland load profile data. The K-means method was also used in this study because of its fast iterative algorithm to sort a large set of data. Upon clustering, the accuracy of the various clusters was verified through the use of the clustering dispersion indicator (CDI). The optimum number of clusters was verified through this way. The clustering was done at two levels, firstly, the monthly load clusters were extracted and secondly, the same procedure was repeated on the annual data.

Using real measurement data, the ant colony search method was applied to the clustering problem [16]. As a benchmark, K-means clustering was initially used in solving the clustering problem. In comparison with the latter method, the ant colony search method produced better results. Furthermore, when the optimal number of clusters was 30, the K-means algorithms failed to cluster the data [16]. The clustering algorithm only yielded empty clusters. This is due to the high number of features the data in question is known to exhibit [16].

The quality of clustering tools are appraised with respect to electricity customer clustering, 234 electricity consumer profiles were used [17]. Clustering results were achieved through tools such as follow-the-leader unsupervised clustering algorithm and the support vector machine (SVM), the former was considered to be the optimal method [17]. Likewise, experimentation by applying various pattern recognition methods: K-means, SOM, fuzzy k-means and hierarchical algorithms are also presented. Six adequacy indicators are examined with respect to verifying the best clustering tools. The K-means and hierarchical clustering algorithms are later proposed.

Similar to [16], the authors in [3] used the K-means method in clustering electricity time series data. The K-means algorithm is a partitioning relocation clustering method (distance-based method) was preferred over the hierarchical, Density-based partition and Grid-Based clustering types. A detail elaboration of the types of clustering methods used in data mining are presented in detail in [13]. The results obtained showed a degree of accuracy in grouping accounts with a similar proportion of energy consumption [3]. The main distinction between the work in [3] and the study in this thesis is that the data set used in [3] was only limited to commercial or industrial customers whereas in addition to commercial and industrial customers, this work also considered residential customers. Again, another dissimilarity is that [3] aims at determining the close and open hours of business by deriving insight from the data.

Fuzzy C-means (FCM) and the Euclidean distance metric was utilized to cluster 25 smart grid users into five clusters [1]. The study in [1] differs from [16], [17], [3] and the study in this thesis in that [1] made used of real-time data. Also, the scope of data considered in this thesis far exceeds the datasets considered in [1] (25 users were considered). A statistical method called fixed mixture model was employed to cluster publicly available data of about 4000 smart meter customers [7]. The research considered the Bayesian information criterion in selecting the ideal number of the clusters to group the users into clusters.

Electricity usage has the tendency to vary for different days even for the same consumer and the other challenge is the high dimensionality of the data. A time-based Markov model was presented to form the dynamics of electricity usage. Also, a symbolic aggression approximation (SAX) was used to reduce the dimensionality of the load curves before the fast search and find of density peaks (CFSFDP) algorithm was applied for customer segmentation [21]. Finally, CFSFDP and adaptive k-means were introduced for the large data set.

Table 1 gives a brief summary of load profiling methods.

Table 1. Summary of the related work.

Techniques	**Methods**	**Number of Clusters**	**Refs.**
Clustering	· K-means, Hierarchical	$[9, 6, 3 \text{ to } 5]$	[3, 19, 20]
Neutral network	· Self organizing maps	$[16]$	[17]
Statistical	· Mixture and Markov models	$[10]$	[7]
Fuzzy approaches	· Fuzzy C-means	$[5]$	[1]
Hybrid methods	· Method combination		[16–18, 21]

The load profiles are grouped such that the profiles with the same behavioural pattern belong to the same cluster and profiles that are different in different clusters. This means the intra-cluster similarities are maximized contrary to the inter-cluster similarities which are minimized. The K-means algorithm does not necessarily find the most optimal configuration corresponding to the global objective function. Also, the algorithm is very sensitive to the initial randomly selected cluster centroids. It is due to this reason that the genetic algorithm is used to optimise the centroids. It complements the K-means algorithm in obtaining good cluster qualities for the load profiling. Despite these downsides, the K-means algorithm is fast, efficient and has been adapted to many domains. It has the capability to handle large data sets and it is easy to implement. And most importantly it gives good results. This is the motivation for using the K-means algorithm in this work. A summary of the K-means algorithm can be found in Table 2.

Table 2. Summary of the K-means algorithm

	Advantages	**Disadvantages**	**Refs**.
K-means	· Simple	· Sensitive to selection of initial centroids.	[22–24]
	· Scalable and Efficient	· Number of cluster has to be defined.	
	· Can handle big data	· Sensitive to noise and outliers.	
	· Linear complexity	· Local optimum solution.	

The K-means algorithm terminates primarily if the data points in the respective clusters are not reassigned to a different cluster or if the maximum number of allowed iterations is attained. For this study, the latter is used.

3 PROPOSED APPROACH FOR ELECTRICITY POWER PROFILING

3.1 K-means clustering

The K-means algorithm is a partition clustering algorithm. It was introduced by J.B. MacQueen in 1967 [25]. The algorithm is based on unsupervised learning used with unlabeled multidimensional data. The goal of the algorithm is to group the unlabeled multidimensional data into K clusters (K is fixed a priori). The K variable represents the number of groups for the partition. It works by iteratively assigning data points to one of the K groups based on the provided features. Each data point is assigned to one unique group. The algorithm is favoured in many application areas such as computer vision, image processing, business analytics etc. Its popularity is due to the simplicity and linear complexity, defined as $O(I * n * K * D)$, where I represents the number of iteration, n is the number of input features, K is the cluster number and D is the dimension of the features [26]. The K-means algorithm includes two steps: 1. Cluster Assignment step 2. Move centroid step. In the cluster assignment step, the idea is to define K centroids for the clusters, one for each cluster. The K-means result is sensitive to the initial centroids, different initial centroid yield different results. Therefore, a good choice is to place them farther away from each other. The next step involves examining each data point and assign the data point to the closest centroid. In the move centroid step, the algorithm calculates the average of all the data points in each cluster and the centroid is moved to that location. This continues until no changes in the clusters occur or until some stopping criterion is met. The algorithm aims at minimizing an objective function, which in this case is the squared error function:

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \|x_i^{(j)} - c_j\|^2 \tag{1}$$

where $k =$ number of clusters,

$n =$ number of data points, and

$\|x_j^{(j)} - c_j\|^2$ is the distance metric used. That is the distance between the load profile $x_j^{(j)}$ and the cluster center c_j. The distance metric used in this case is the Euclidean distance. The Euclidean distance formula is given in Equation 2.

$$d(x, c) = \sqrt{\sum_{k=1}^{x} (x_k - c_k)^2}. \tag{2}$$

Assuming X is the set of load profiles with $X = x_1, x_2, ..., x_n$ and $V = v_1, v_2, ..., v_k$ is the set of cluster centroids, the K-means algorithm proceeds by the following steps:

1. c cluster centroids are randomly selected.
2. The distances between each load profile and the cluster centroids are calculated.
3. Assign the load profile to a cluster centroid with the minimum distance.
4. Recalculate the new cluster centroids as follows:

$$V_i = \frac{1}{|c_i|} \sum_{j=1}^{c_i} x_i \tag{3}$$

   c_i represents the number of data points in the i^{th} cluster.

5. Recalculate the distance between each load profile and the new cluster centroids.
6. If no single load profile is reassigned to a cluster centroid ,the algorithm stops else proceed to step 3.

3.2 Genetic algorithm

Genetic algorithms are biologically-inspired heuristic search optimization algorithm. They are inspired by Charles Darwin's theory of evolution i.e the survival of the fittest. The algorithm exhibit the process of natural selection in which the fittest individuals are chosen to reproduce the offspring of the next generation. The genetic algorithm essentially replicate the way in which life uses evolution to find solutions to real world problems [27].

There are five phases considered in a genetic algorithm [27]:

1. Initial population
2. Fitness function
3. Selection
4. Crossover
5. Mutation

A brief description of these phases is given below:

Initial population

Population of randomly generated solutions to the problem. Clearly, randomly generated solutions to the problem might not be too ideal.

Fitness

The fitness quantitatively evaluates how fit a given solution is or how fit individuals can be produced from the given solutions i.e., the fitness ability of an individual to compete with others. A fitness score is assigned to each individual, the selection of an individual for reproduction depends on its fitness score [27].

Selection

In the strive to achieve convergence, the best offsprings are selected as parents in the new parental population. The selection of the offsprings are based on their fitness values [27].

Crossover

During crossover the genetic material of the parents are combined. This can be thought of as mimicking the mating process in real life. By combining certain traits from two or more individuals, the hope is that a 'fitter' offspring will evolve with the best traits inherited from the parents [27].

Mutation

Mutation-operators provides random changes to the population by disturbing them. Mutation typically allow very small changes at random to the individual genomes [27]. Mutation maintains diversity within the population and help prevent fast convergence.

Termination

When convergence is attain the algorithm terminates. At this point it can be said that the algorithm has provided a solution to the problem.

The GA cycle is given in Figure 3.

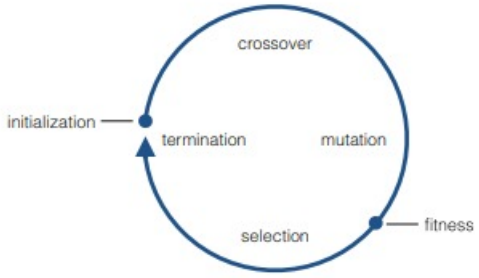


Figure 3. Genetic Algorithm cycle [27].

4 EXPERIMENTS AND RESULTS

The proposed algorithms are implemented on Matlab R2017a version, on a Windows 10 machine with 8GB of RAM. The Matlab inbuilt function 'Kmeans' and 'ga' were used for the implementation. The software provides flexibility in reading and displaying stored files.

4.1 Description of data

The data set used in the experiment are time series hourly electricity consumption data for 13601 households in Southern Finland. The data are based on hourly loads recorded for a span of one year. The rows in the raw data set represent the time-stamps and the columns the respective customers. The dimension of the load data is 8760×13601.

4.2 Pre-processing

The given data set was pre-processed to remove missing values. In checking the load data for missing information the Matlab inbuilt functions 'isnan' was used. A single user's data was found with missing information. Only that particular user was excluded from the final data set used for the experiments. The K-means algorithm, in this case, uses the Euclidean distance metric. The Euclidean distance is known to be biased due to the scale of the measurements, to this regard the raw electricity consumption data was standardized, so that it has zero mean and unit variance. The following formula was used for the standardization:

$$\bar{X} = \frac{(X_i - \mu_i)}{\sigma_i}. \tag{4}$$

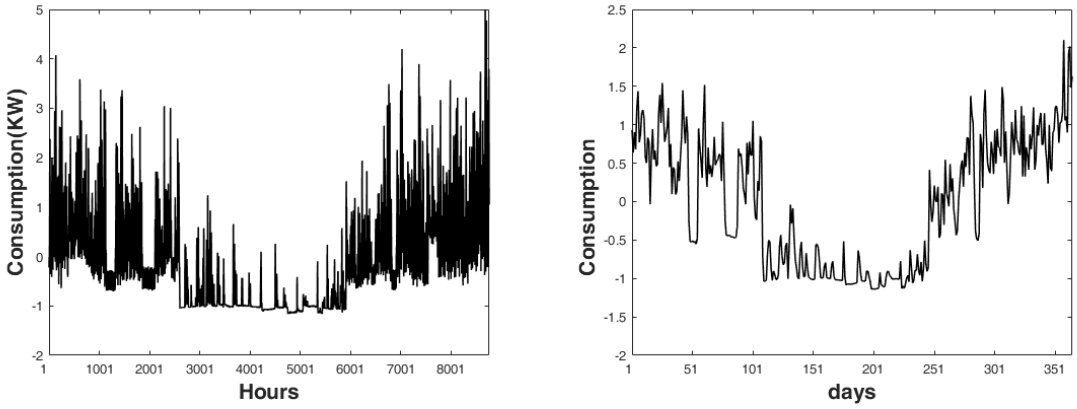
where X is the load profile data, μ is the mean and σ represent the variance. For each respective user load, μ_i and σ_i represent the mean and variance respectively of the entire data for that particular user.

4.3 Dimensionality reduction

As the number of dimensions increases, the distance between any two points in the same data sets converges (the maximum distance and the minimum distance between any two points will be identical) [28]. This tends to be an issue with the Euclidean distance metric. Reducing the dimensionality prior to the K-means clustering can alleviate this problem and considerably help with the computation. The dimensionality reduction technique used was adopted from [29].

To reduce the load profile data from n dimensions to N dimensions, the data was divided into z equally-sized frames. The mean of all the data within this frame was computed and a vector N of all the mean values derived becomes the new representation of the original data. This dimension reduction was needed only for deriving the annual profiles. The whole data set was considered in building the annual profiles hence, the need to compress the size of the data.

For this study, the data was divided into 24 equally-sized frames (8760 rows into 24 equal-sized frames), 24 because each user provides 24 data points a day. In simple terms, the average of the load data provided in a day represents the electricity consumption on that particular day. For the annual load profiles, the dimension of the data is reduced from 8760×13600 to 365×13600. Figure 4 shows the full load profiles and the corresponding dimensional reduced profiles. The dimensionality reduced profile was obtained by the method describe above. It can be seen that the shape of the two profiles has not changed.



(a) (b)

Figure 4. Load profile of residential flats: (a) A full load profile. (b) Dimensionality reduced load profile.

4.4 Evaluation criteria

Various methods can be used to quantify the performance of a clustering algorithm as well as to provide a technique for the selection of the appropriate number of clusters. The evaluation criteria can be categorized as similarity-oriented and classification-oriented [30].

In determining the optimal number of clusters, three validity indices Silhouette, David-Bouldin and Calinski-Harabasz index were used:

4.4.1 Silhouette index

In silhouette analysis, the separation distance between the clusters is studied. It gives a measure of closeness between the points in one cluster to the points in the neighbouring clusters. The formal definition of this quality index was adopted from [31].

Let $X = x_1, ..., x_N$ be the load profile data set and let $C = c_1, ..., c_k$ be its clustering in some k clusters. Let us denote $d(x_k, x_i)$ to be the distance between x_k and x_i. Let $c_j = x_i^j, ..., x_{m_j}^j$ be the j^{th} cluster where $j = 1, ..., k$ and $m_j = |c_j|$. a_i^j denotes the average distance between the i^{th} vector in the cluster c_j and the vectors in the same cluster. The average distance a_i^j is hence given by :

$$a_i^j = \frac{1}{m_j - 1} \sum_{k=1, k \neq i}^{m_j} d(x_i^j, x_k^j),\ i = 1, ..., m_j. \tag{5}$$

The minimum average distance between the i^{th} vector c_j and all the vectors clustered in cluster c_k , where $k = 1, ..., K$ and $k \neq j$ is given as follows :

$$b_i^j = \min_{n=1,...,k, n \neq j} \left(\frac{1}{m_n} \sum_{k=1}^{m_n} d(x_i^j, x_k^n) \right), i = 1, ..., m_j. \tag{6}$$

The i^{th} vector silhouette width in cluster c_j is given below:

$$s_i^j = \frac{b_i^j - a_i^j}{\max(b_i^j, a_i^j)}. \tag{7}$$

The silhouette width is in the range $[-1, 1]$. The silhouette of a cluster c_j given as:

$$s_j = \frac{1}{m_j} \sum_{i=1}^{m_j} s_i^j. \tag{8}$$

The algorithm for determining the optimal number of clusters using the Silhouette index is given as follows:

1. Perform K-means clustering for the range of values of K.
2. For each value in the range, an average Silhouette was calculated for the observation.
3. A plot of the curve according to the average silhouette was generated.
4. The location of the maximum is the optimal number of clusters.

The Matlab inbuilt function 'evalclusters' was used to achieve this.

4.4.2 Davies–Bouldin index

The Davies–Bouldin (DB) index was introduced in 1979 by David L. Davies and Donald W. Bouldin. It is the ratio between the within-cluster distances and the between-cluster distances and computing the average over all clusters [31].

The formal definition of the Davies-Bouldin index was adopted from [32]. Let δ_k denote the mean distance of the point in the cluster c_k to their centroids G^k:

$$\delta_k = \frac{1}{n_k} \sum_{i=1}^{n_k} \|M_i^k - G^k\|. \tag{9}$$

where M_i^k is the n-dimensional feature vector assigned to cluster c_k, and n_k is the size of the cluster. Let us denote also,

$$\Delta_{kk'} = d(G^k, G^{k'}) = ||G^k - G^{k'}||. \tag{10}$$

that is, the distance between the centroid G^k and $G^{k'}$ of clusters c_k and $c_{k'}$. For all indices $k' \neq k$, the Davies-Bouldin index is as follows:

$$C = \frac{1}{K}\sum_{i=1}^{K} \max_{k' \neq k}\left(\frac{\delta_k + \delta_k'}{\Delta_{kk'}}\right), \text{where } K = \text{number of clusters.} \tag{11}$$

The algorithm to find the optimal value using the Davies-Bouldin index is similar to the Silhouette method. The algorithm is given below as:

1. Perform K-means clustering for the range of values of K
2. For the values in this range, calculate the Davies-Bouldin index.
3. A plot is generated for each value of K.
4. The location of the minimum is considered to be the optimal number of clusters.

4.4.3 Calinski-Harabasz index

In the Calinski-Harabasz index, the comparison of the between-clusters variance to the within-cluster variance is made. The index was first introduced in 1974. The formal definition is derived from [32] and it is given as:

$$C = \frac{BGSS}{WGSS} \times \frac{N-K}{K-1} \tag{12}$$

where K represent the number of clusters, N is the total number of load profiles. The overall within-cluster variance is denoted $WGSS$ and the overall between-cluster variance as $BGSS$ [32].

$BGSS$ is calculated as the total sum of squares subtracted from $WGSS$. The total sum of squares is the squared distance of all the load profiles from the centroids.

The steps of the algorithm is given as follows:

1. K-means clustering is computed for values in the range K.
2. Derive the Calinski-Harabasz index for each value in the range.
3. Plot of the curve according to the derived Calinski-Harabasz index.
4. From the plot, the location of the maximum is the optimum value.

4.5 Implementation of experiments

The experiment was divided into two stages. In the first stage, the clustering was performed using the K-means method. For studying the annual weekday profiles, the weekend loads were separated from the data before the clustering process. The motivation for doing this was to study the electricity consumption pattern during the weekdays compared to the weekends. It is expected that the load curves in the weekdays should vary from the ones of the weekends for residential households. This is evident by the fact that many residents should be away during working hours hence imposing some limitations on electricity consumption. The clusters are studied for finding evidence of this.

When using the Matlab inbuilt function 'Kmeans', the function's default parameters were used. The only change made was to the 'maximum number of iteration (MaxIter)' which was set to 1000 iterations (this was to provide more running time for the genetic algorithm in the optimisation step). The load data for the clustering algorithm and the value for K was also supplied. The outputs are the clusters computed for the load profiles and the cluster centroids. To determine the value for the optimum number of clusters, a range of values for K were considered. These values initially range from 2 to 20. The optimal number of clusters was identified by using the Silhouette, David-Bouldin and Calinski-Harabasz indices recorded for each of the value of K.

After the optimal number of clusters is known, the second stage of the experiment involves the optimization of the centroids corresponding to the optimal K . The goal is to find better cluster centroids. The Matlab inbuilt function 'ga' was used in the optimization. The steps of the algorithm are given as follows:

1 A random selection of $K \times d$ load points are chosen from the load data. K is the number of clusters and d represents the number of loads data.

2 Let c represent the cluster centroids and $x = K \times d$ which is the random selection of the data points.

3 The fitness function in this case the Euclidean distance, minimizes the distance between c and x.

4 To improve the quality of the clusters, the K-means clustering is repeated for the second time. The output after the optimisation with GA now serves as the initial set of centroids for the K-means algorithm. In short, the K-means algorithm was run using random initial centroids (chosen from the data). The centroids provided by the

K-means algorithm was optimised using GA. After the optimisation, the K-means algorithm is rerun to cluster the data but using the output from the optimisation as initial centroids.

The methods are known as 'K-means' which is the standard K-means algorithm and 'K-means with GA' which involves the standard K-means but with optimised initial centroids by GA to achieve improved cluster quality.

4.6 Results

4.6.1 Selecting the number of clusters for the annual load profiles

During finding the appropriate number of clusters, values ranging from 2 to 20 were tested. The validity indices for each of these values are derived. In Figure 5 and Table 3, the Silhouette, Davies-Bouldin and Calinski-Harabasz index indicates that the appropriate number of clusters is 2 for the annual load profiles.

Table 3. Results of the three validity indices. Bold numbers denote the best values for the indices.

No. of clusters	Silhouette	Davies–Bouldin	Calinski–Harabasz.
2	**0.2151**	**2.5785**	**1594.8**
3	0.1181	3.2887	1154.4
4	0.1136	3.0766	945.2
5	0.0922	3.5353	787.0
6	0.0756	3.4909	693.8
7	0.0748	3.3371	632.9
8	0.0747	3.2770	571.7
9	0.0615	3.3888	523.04
10	0.0577	3.3862	479.7
11	0.0631	3.3729	445.9
12	0.0568	3.2304	405.4
13	0.0587	3.4175	392.0
14	0.0514	3.6137	368.2
15	0.0563	3.1679	336.1
16	0.0501	3.5583	331.6
17	0.0511	3.5669	316.9
18	0.0451	3.6810	302.1
19	0.0424	3.7351	286.3
20	0.0427	3.8034	275.6

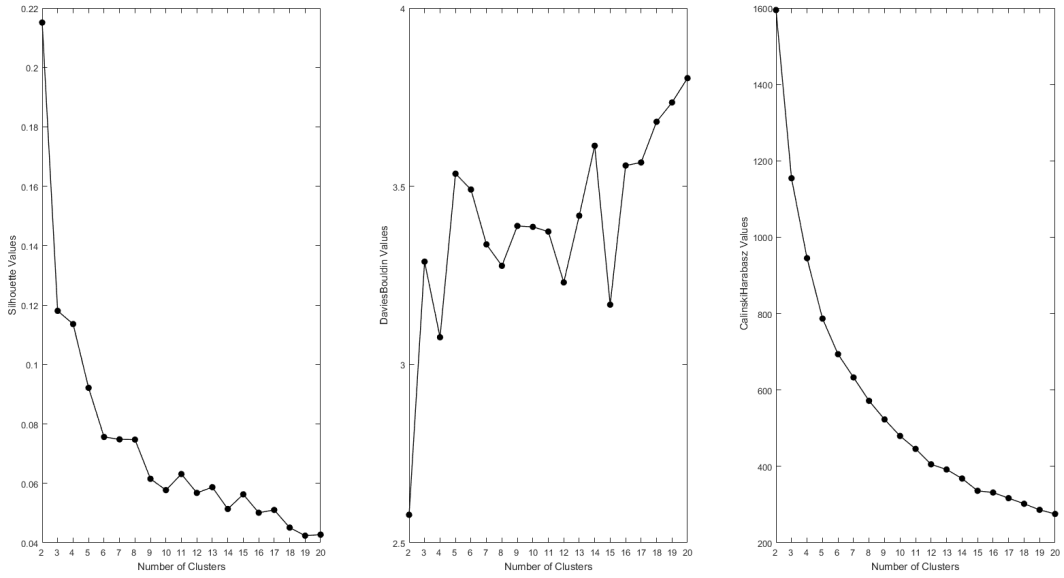


Figure 5. Silhoutte, Davies-Bouldin and Calinski-Harabasz index.

The two clusters for the annual profiles are shown in Figure 6.

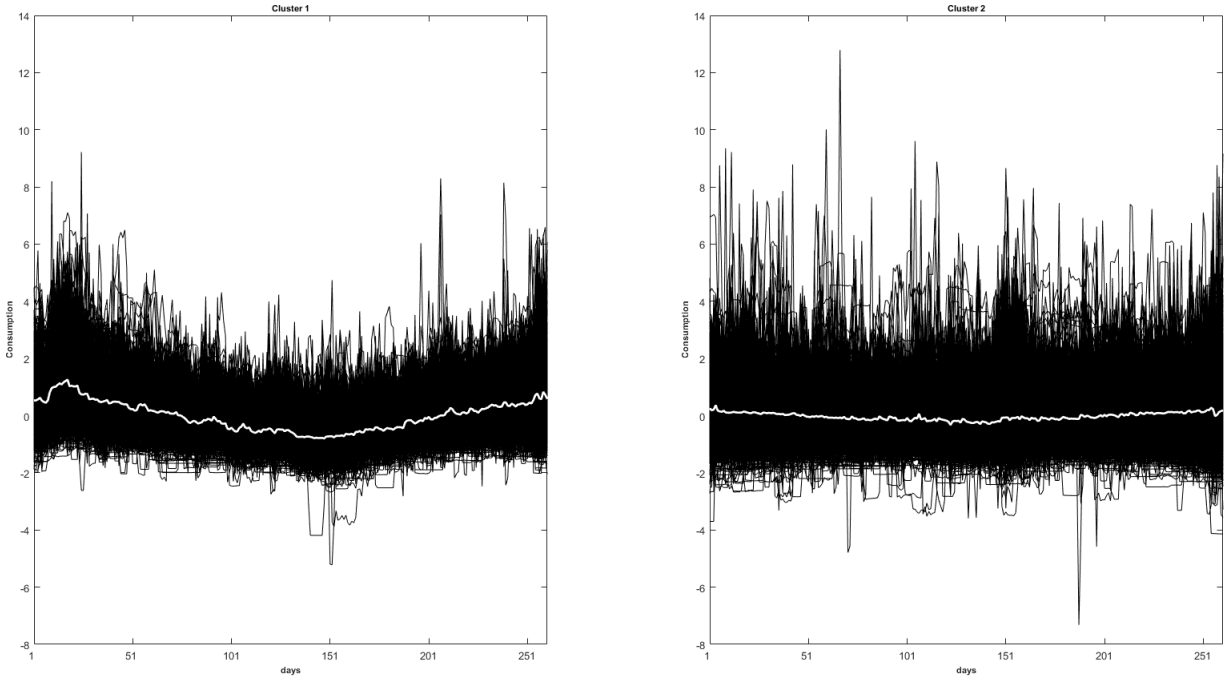


Figure 6. Annual weekday load profiles, 3901 and 9699 load profiles in each respective cluster.

4.6.2 Selecting the number of clusters for the daily load profiles

Similarly, the Silhouette, Davies-Bouldin and Calinski-Harabasz index indicates that the appropriate number of clusters for the daily load profiles is also 2. The values of the three validity indices are represented in Table 4. Figure 7 contains the plot of these values.

Table 4. Results of the three validity indices for the daily weekday load profiles. Bold numbers denote the best values for the indices.

No. of clusters	Silhouette	Davies–Bouldin	Calinski–Harabasz.
2	**0.4461**	**1.2454**	**7077.1**
3	0.3555	1.4624	5603.0
4	0.3548	1.4032	4982.8
5	0.3326	1.3316	4591.6
6	0.3511	1.2797	4235.2
7	0.3282	1.3276	3968.7
8	0.3219	1.3375	3722.9
9	0.2908	1.3646	3497.2
10	0.2968	1.3701	3290.7
11	0.2936	1.3667	3119.5
12	0.2832	1.3751	2989.2
13	0.2745	1.3937	2838.8
14	0.2892	1.3666	2750.3
15	0.2864	1.3895	2649.7
16	0.2700	1.3904	2553.0
17	0.2671	1.3752	2473.6
18	0.2631	1.4035	2401.7
19	0.2696	1.3866	2266.1
20	0.2525	1.4264	2266.9

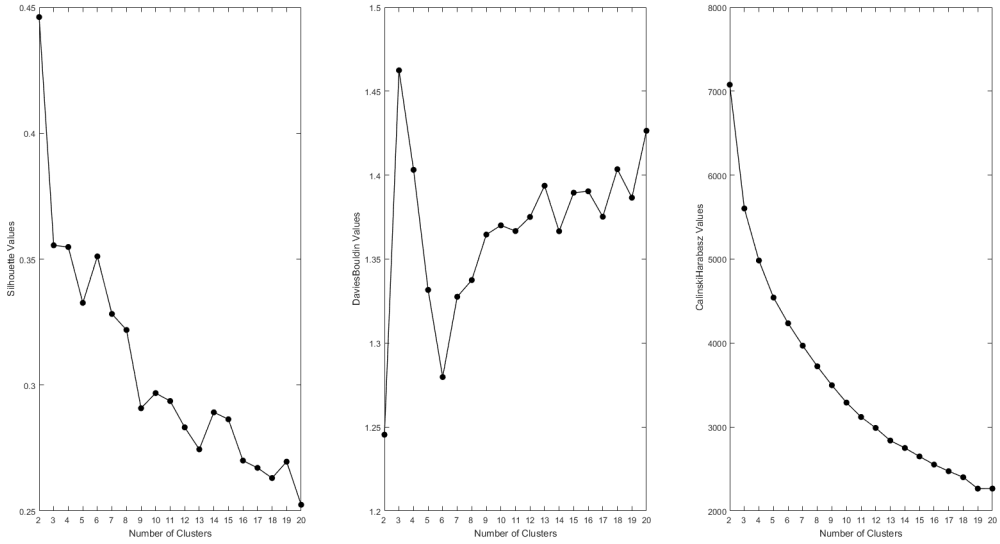


Figure 7. Silhoutte, Davies-Bouldin and Calinski-Harabasz index.

In Figure 8 the two daily weekday load profiles are represented.

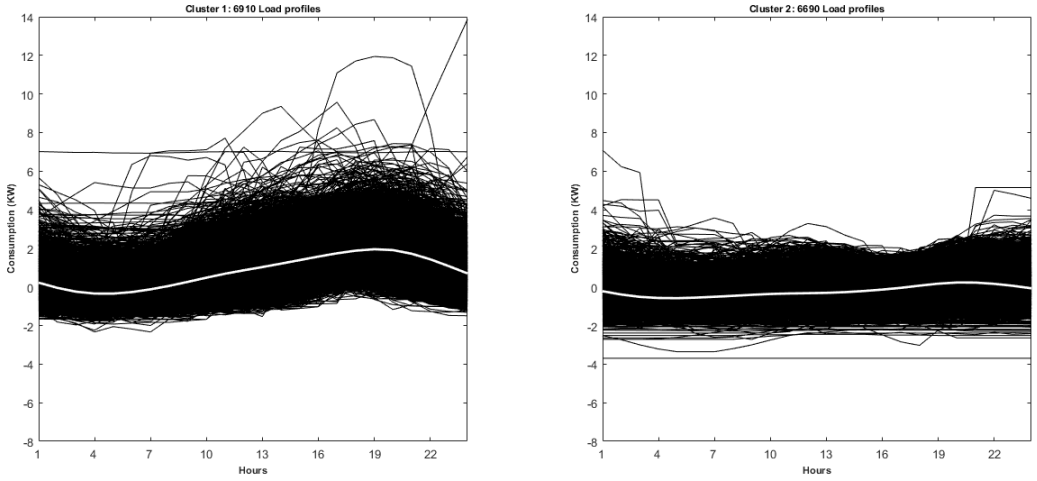


Figure 8. Daily weekday load profiles, 6910 and 6690 load profiles in each respective cluster.

Figure 6 and 8 present the new clusters for the annual weekday profiles and the daily weekday profiles respectively. From the Figures, the clusters seem to show only two groups of users i.e. users with a flat electricity consumption pattern and those users whose electricity consumption varies across the year or day. An observation of the formed clusters showed that a lot of averaging occurred and some potential unique traits of the respective users are not exhibited. Also, considering the number of users in the data, 2 clusters is too small to represent the electricity consumption behaviours of these users. Previous work considered in this study had the optimum number of clusters higher that 2 as seen in Table 1. Therefore, it is safe to argue that 2 is not appropriate for the categorization of the load profiles. Other ways of choosing the appropriate number of clusters were examined. A different range of values needs to be considered for the appropriate number of clusters. The Davies-Bouldin index is chosen because the other two validity indices, in this case, do not seem to be the appropriate method for selecting the number of clusters. The Davies-Bouldin index values that seem to be the potential solutions are looked at, these values are 4, 8, 12 and 15 for the annual weekday load profiles and 6 and 14 for the daily weekday profiles. Before the potential Davies-Bouldin values are analysed in detail ,we first study the the within-cluster sum of squares for both the daily and annual load profiles

4.6.3 The within-cluster sum of squares for annual load profiles

To further analyze the appropriate number of clusters, the within-cluster sum of squares was applied. The degree of variability of the load profiles in each cluster is given by the within-cluster sum of squares (WCSS). The WCSS decreases as the number of clusters increase. The appropriate number of clusters can be selected this way, the hint is to choose the number of clusters from which the WCSS drop is not very large.

In Figure 9, the WCSS drop is not large around 8 therefore, the annual load profiles can be categorized into 8 clusters.

The WCSS is supposed to decrease and stay low as the number of clusters increase. The situation is different when the number of clusters is 12 and 15, the WCSS increased instead of staying low. The percentage of variance as a function of the number of clusters is looked at. A number of clusters should be chosen so that an addition of another cluster does not give a better modelling of the data. In this view, even though the WCSS drop did not stay low at 12 and 15, the two values do not give a much better result than when the number of clusters was 8.

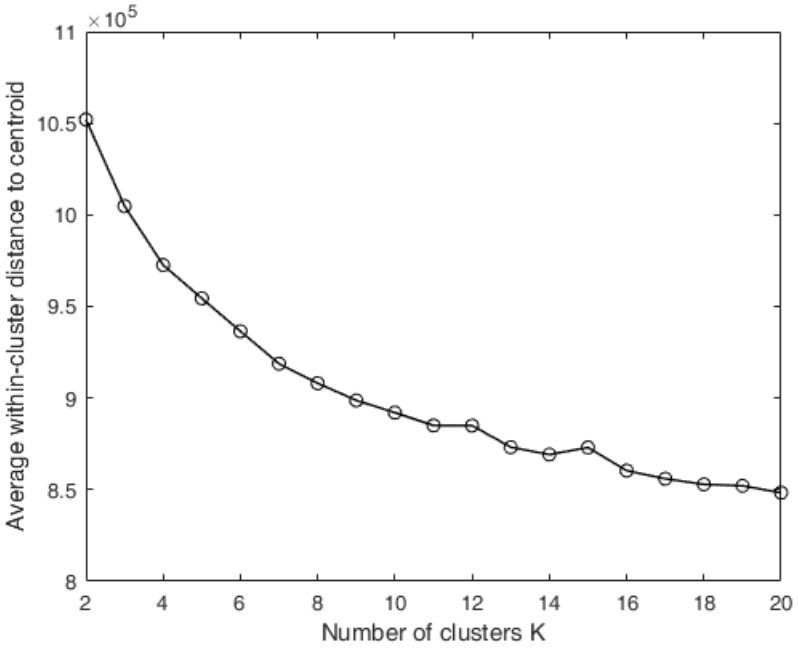


Figure 9. Within-cluster sum of squares for annual load profiles.

4.6.4 The within cluster sum of squares for daily profiles

The within-cluster sum of squares is also utilized to study the appropriate number of clusters for categorizing the daily load profiles. Figure 10 provides the WCSS plot.

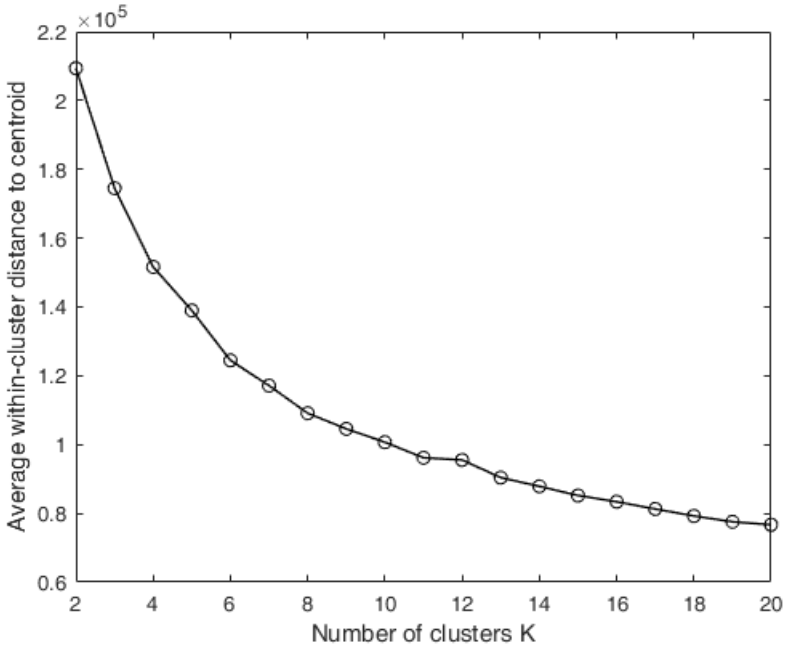


Figure 10. Within-cluster sum of squares for daily load profile.

From the plot in Figure 10 it is seen that the WCSS drop is not substantial around cluster 6 and 7. These values can, therefore, suggest the appropriate number of clusters. The WCSS did not stay low for all the values analyzed. At 12 the WCSS increased instead of dropping. The same argument used with the annual load profiles also applies here. The value at 12 is still lower than the value at the appropriate number of clusters.

4.7 The cluster representation for the annual load profiles

For each of the cases considered, i.e., Case 1, Case 2, Case 3 and Case 4 (Corresponding to 4, 8, 12 and 15 clusters respectively), the number of profiles in each case is given in Table 5.

Table 5. The number of consumers in each of the cases considered above.

Cluster	**Case 1 (4 profiles)**	**Case 2 (8 profiles)**	**Case 3 (12 profiles)**	**Case 4 (15 profiles)**.
1	1538	942	850	794
2	2166	1068	497	466
3	4610	3809	525	2142
4	5286	1244	569	410
5	•	2304	1477	460
6	•	850	289	242
7	•	742	487	437
8	•	2641	1811	683
9	•	•	569	1619
10	•	•	2659	1799
11	•	•	1384	1235
12	•	•	2483	1481
13	•	•	•	385
14	•	•	•	1044
15	•	•	•	403

Case 1: 4 **Annual weekday profiles**

In Cluster 1 found in Figure 11, a peak appeared in the first month of the year. The high electricity consumption rate declined as the year proceeds, this trend continued until mid-year. Generally, the weather in Finland is friendlier around this time of the year hence electric heating is not a necessity, evident in the relatively stable electricity consumption showed. Around the end of the year, the consumption rates are shown to be high once again, this rise in the pattern of consumption is attributed to the drop in temperature which is experience around the beginning of the winter season. Another peak in electricity consumption was observed at this time. This profile represent residential homes where district heating is not provided so resident have to result to providing heating for themselves during the winter period. Some confidence can be given to this claim due to the major peaks recorded around the time when the temperatures are very low.

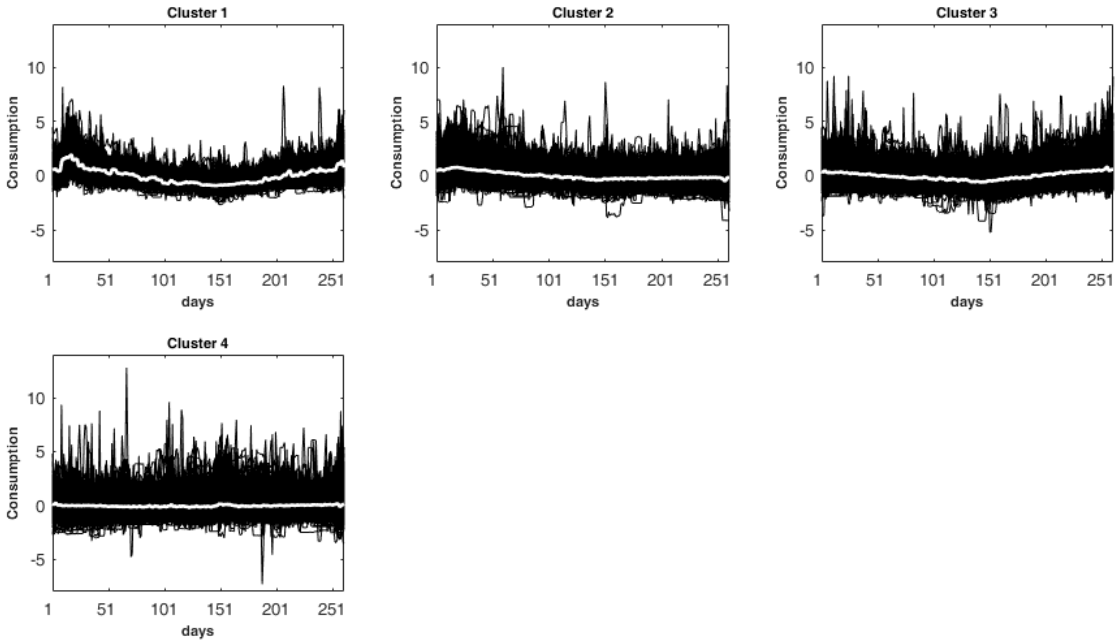


Figure 11. Case 1: 4 Annual weekday profiles.

In Cluster 1 found in Figure 11, a peak appeared in the first month of the year. The high electricity consumption rate declined as the year proceeds, this trend continued until mid-year. Generally, the weather in Finland is friendlier around this time of the year hence electric heating is not a necessity, evident in the relatively stable electricity consumption showed. Around the end of the year, the consumption rates are shown to be high once again, this rise in the pattern of consumption is attributed to the drop in temperature

which is experience around the beginning of the winter season. Another peak in electricity consumption was observed at this time. This profile represent residential homes where district heating is not provided so resident have to result to providing heating for themselves during the winter period. Some confidence can be given to this claim due to the major peaks recorded around the time when the temperatures are very low.

In Cluster 2 of Figure 11, the electricity consumption during the beginning of the year showed a stable behaviour, but a peak in the consumption rates was recorded during the middle of the first month (roughly around the 15^{th}). This rise in consumption corresponds to the beginning of the year when temperatures are at their lowest. The consumption rate became stable for the rest of the year. The consumption rate takes up again around the end of the year when little peaks are seen to appear. These are most probably houses with electrical heating. The difference between this profile and the first profile in Figure 11 is that the consumption rate showed a more stable behavior.

In Cluster 3 of Figure 11, a low but noticeable peak appeared around January after which the behaviour of the profiles recorded a stable but downwards trend, this pattern stays consistent. Around June/July, the consumption rate declined even further. At this time, the days are normally warmer compared to the rest of the year. A stable increase is recorded right afterwards. The customers in this profile can be attributed to residential homes were a form of air conditioning or cooling system is absent. The absent of a cooling system is noticeable in the downward trend in consumption rates exhibited for warmer days.

The profiles in Cluster 4 of Figure 11 recorded a stable consumption pattern in electricity usage in the beginning of the year. A rise in the electricity consumption begins shortly afterwards, evident in the peaks seen at roughly around June. This profile is attributed to residential homes in which some form of cooling system is present. This should explain the peaks around that time of the year when temperatures are normally not low. The major peaks declined at around July/August. The pattern stays almost the way through the rest of the year. As the year elapsed the consumption rates began to increase again.

Case 2: 8 **Annual weekday profiles**

Cluster 1, 2, 4 and 5 of Figure 12 show similar consumption behaviour to profiles in Figure 11 i.e, Cluster 1, 2, 3 and 4 respectively.

Cluster 3 of Figure 12 showed the same pattern of energy consumption throughout the

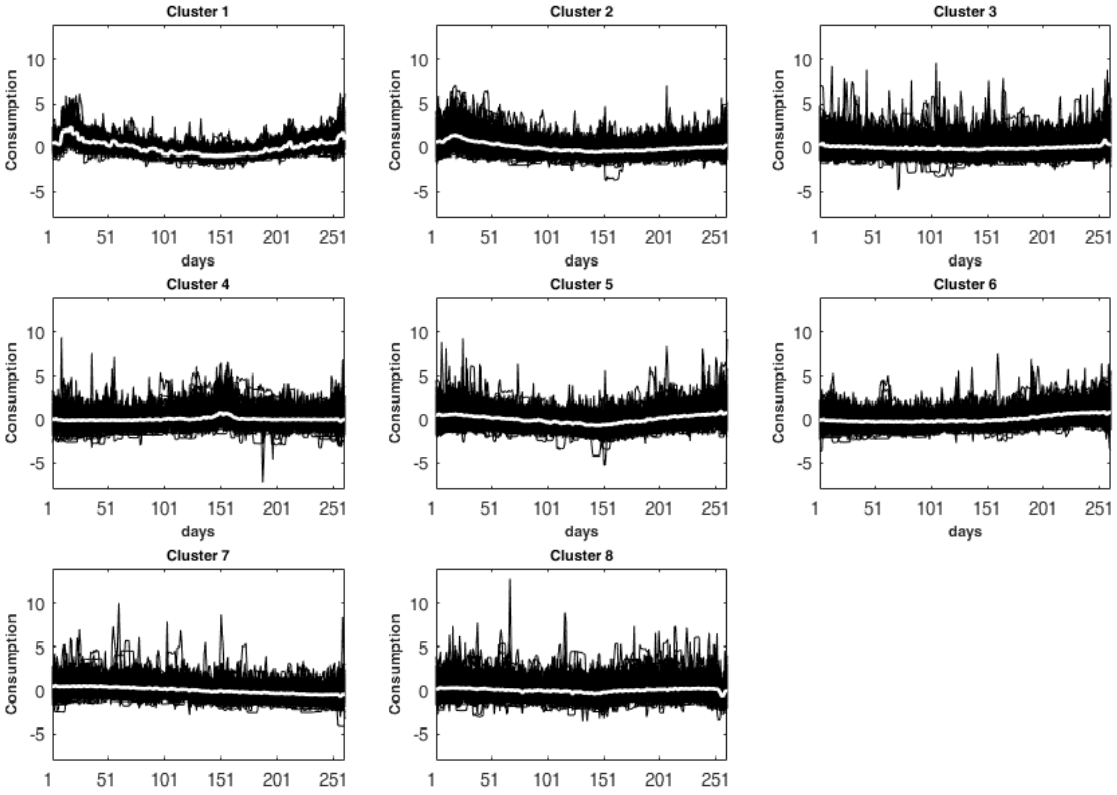


Figure 12. Case 2: 8 Annual weekday profiles.

year. An increase in the electricity consumption is seen in the form of peaks during the early months of the year, but this phenomenon did not last long as the constant pattern continued. Besides few individual peaks, a major peak also occurred at the end of the year.

In Cluster 6 of Figure 12, a rise in the energy consumption is recorded after June. In this profile, as the end of the year approach, the rate of electricity consumption is noticeably seen to be on the rise. Also, some major peaks formed at the latter end of the year.

In Cluster 7 of Figure 12, the consumption pattern showed almost the same behaviour as Cluster 3 of the same figure. The only noticeable difference is that as the year progressed a decline in the energy consumption is seen in Cluster 7. This trend stays consistent for the whole year. Besides few individual peaks, no major peaks can be seen. These profiles can be attributed to load curves produced by industrial or commercials customers.

The consumption of electricity by users in Cluster 8 of Figure 12 began very stable. Despite few individual peaks shown by different users. The energy consumption behaviour remained very much the same until around May. A steady rise in energy usage is noticed around June. This trended until the end of the year. The consumption rate slightly

declined at some point, but quickly took up again as the year ends.

Case 3: 12 Annual weekday profiles

In Figure 13, some of the clusters formed have already been seen and addressed previously, i.e., in Case 1 and 2. Some new profiles different from those found have also been created. The profiles show some similarity visually, for instance, Cluster 8, 9 and 11. The most noticeable characteristic exhibited by these clusters is the decline in the energy consumption rates as year advances.

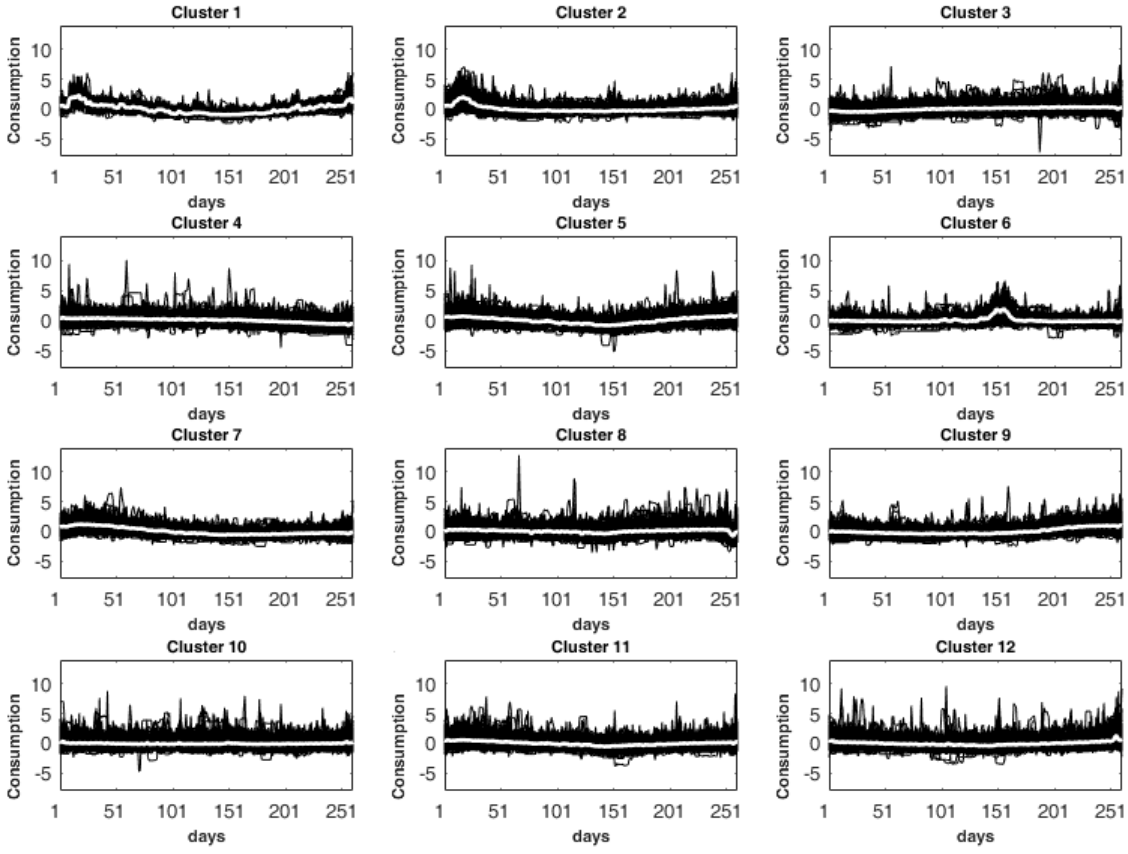


Figure 13. Case 3: 12 Annual weekday profiles.

A minor rise in the electricity consumption is observed in Cluster 7 of Figure 13. Although these peaks are very much noticeable, they are not major peaks. The peaks stayed constant until February, but promptly declined afterwards. At roughly around June, the consumption pattern begins to rise gently. This steady rise is noted throughout the year. This profile can be associated with homes where district heating is absent as evident in the

rise and decline of electricity consumption, i.e., energy is demanded more in the colder months and less in the warmer seasons. Cluster 2 and Cluster 7 have a similar pattern of electricity consumption, However, the peaks in the beginning of the year for Cluster 2 are stronger and more prominent.

Case 4: 15 **Annual weekday profiles**

Like Figure 13, some of the profiles in Figure 14 have already been met. Again, after Case 2, the resulting clusters of the other cases are not very distinct in many ways. In

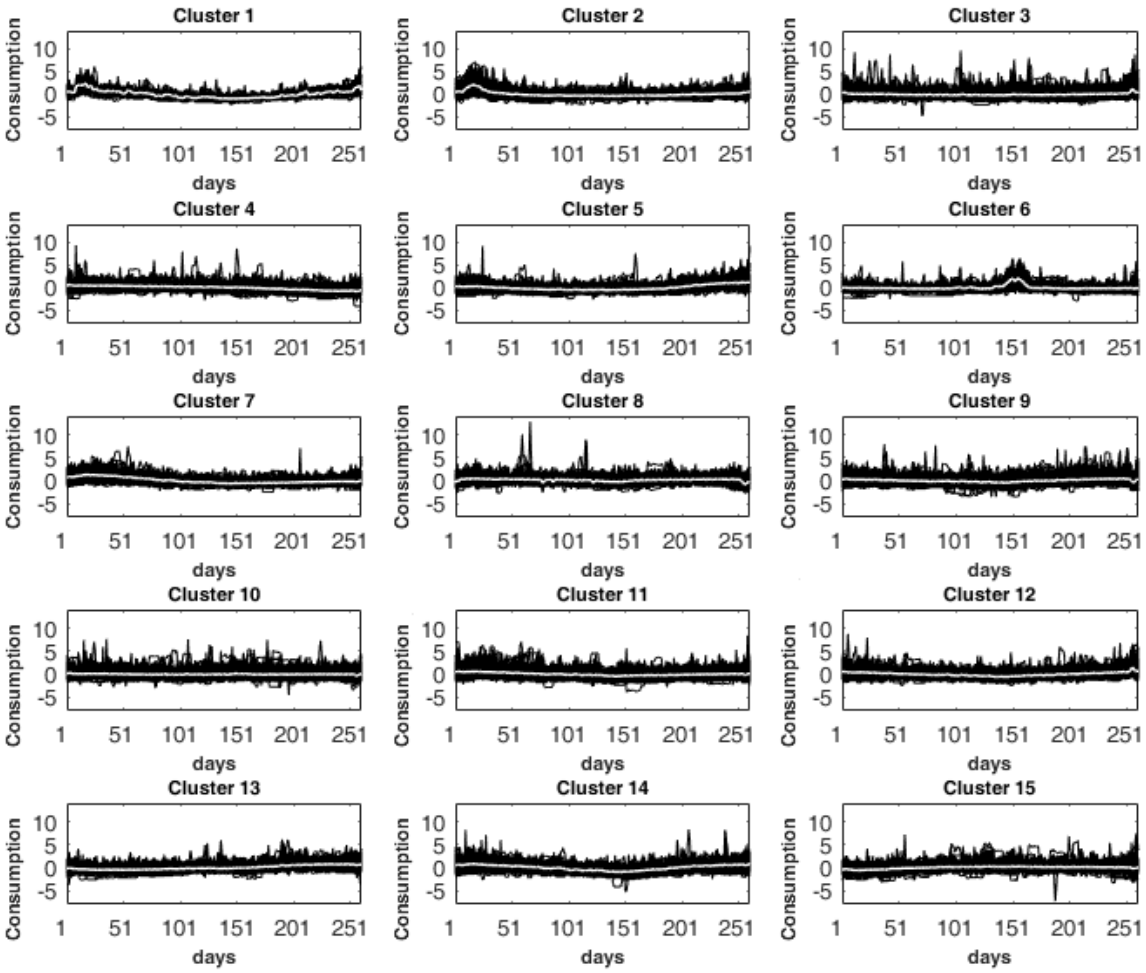


Figure 14. Case 4: 15 Annual weekday profiles.

Figure 14, Cluster 8, 9, 10, 11 and 15 are very related and perhaps these profiles should not form individual clusters but rather be together. The same implies to cluster 5 and 13.

4.8 The cluster representation for the daily load profiles

For each of the cases considered for the daily load profiles i.e. Case 1 and Case 2 (Corresponding to 6 and 14 clusters respectively), the number of profiles in each case is given in Table 6.

Table 6. The number of users in each of the cases considered for the daily load profiles.

Cluster	**Case 1 (6 profiles)**	**Case 2 (14 profiles)**.
1	2333	881
2	4150	1990
3	1499	377
4	2390	1142
5	1059	374
6	2169	1373
7	•	1308
8	•	431
9	•	1005
10	•	893
11	•	1660
12	•	1340
13	•	396
14	•	430

Case 1: 6 **daily weekday profiles**

In Cluster 1 of Figure 15, during the early hours of the day, the need for electricity remains moderately low. At about $6:00$ the demand for energy increased. This trend continued until $14:00$ when finally the demand subsided. This decrease in electricity consumption continued throughout the day. In this profile, the energy requirements were high in the mornings and a greater part of the afternoon when the demand started to decline. This can be attributed to homes were residents spend a good part of the mornings and afternoons at home. This profile might indicate households were residents are away from their homes at mid-day, perhaps they work in the evenings and nights. It can also indicate profiles for restaurants and small cafeterias. A closer look at the profile supports this claim.

The pattern of energy consumption is somewhat higher at $1:00$ in Cluster 2 of Figure 15 compared to the other hours (in the same profile). A reduction in the electricity usage is evident from around $5:00$, this steady decrease trended for a couple of hours until $9:00$

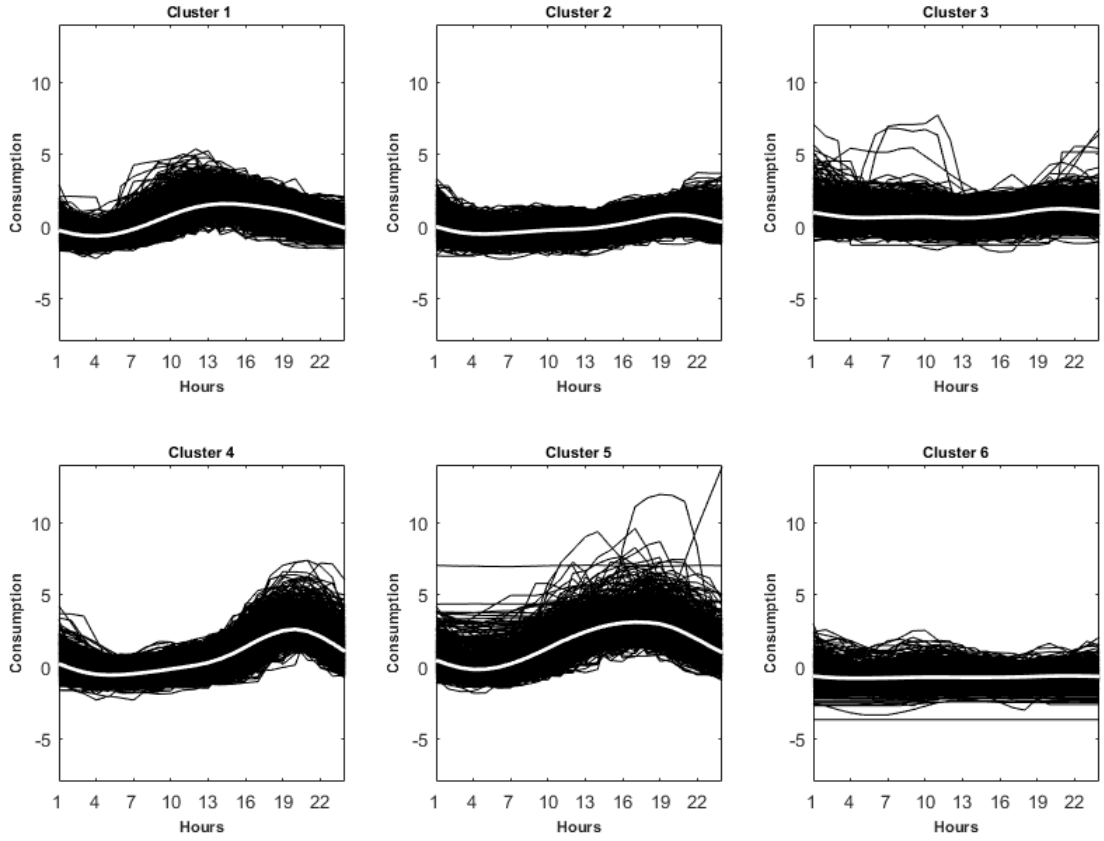


Figure 15. Case 1: 6 daily weekday profiles.

when a slight rise is noticed. From $12:00$ to $16:00$ the consumption rate was steady. No major peaks are detected. Starting from $17:00$ the demand for electricity increased, but at a steady rate. This pattern remained until midnight.

Cluster 3 and Cluster 6 of Figure 15 are comparable in some ways. The only obvious difference is that instead of the profile showing almost identical pattern of power consumption the entire day, the profiles in Cluster 3 showed some upward and downward trend.

In Cluster 4 of Figure 15, shortly after $7:00$ a greater need for energy was registered by the users in the profile. This trended during the mornings, afternoons and the evenings. This profile is distinct from the previous profiles in that the peaks occur at around $19:00$. At $20:00$ the demand for electricity for the first time in many hours began to drop. This fall in power consumption corresponds to the bedtime of many working people. The pattern of electricity consumption is identified to be on the rise throughout the day, this can indicate households were the devices are left switched on even when not in use.

The early part of the morning manifested a low energy consumption rates for the users

of Cluster 5 of Figure 15. By $7:00$, the energy requirement of the users increased. The need for energy increased relatively with every passing hour. At $14:00$ the consumption rate peaked. This plateaued towards $16:00$ when demands began to fall. There was no other significant need for energy onward. Users in this profile showed an urgent need for electricity in the afternoons compared to the night time. The users majorly showed a decline in energy at $19:00$ and this trend lasted for the remainder of the day.

In Cluster 6 of Figure 15, the power consumption remained relatively the same for the whole day. Besides few outliers, the pattern of consumption stays almost the identical. This profile can be attributed to users with a very stable power consumption behaviour. This profile can indicate industrial or commercial users.

Case 2: 14 **daily weekday profiles**

In Figure 16, the electricity consumption nature of some of the user profiles are alike. The profiles vary only in the time of the day the demand for energy peaked. For example, In Cluster 4 and 5 of Figure 16, the shape of the two user profiles are similar, however, the daily peaks occurred at a different time in both profiles. The same refers to Cluster 11 and 9 as well as Cluster 8 and 10 of the same Figure.

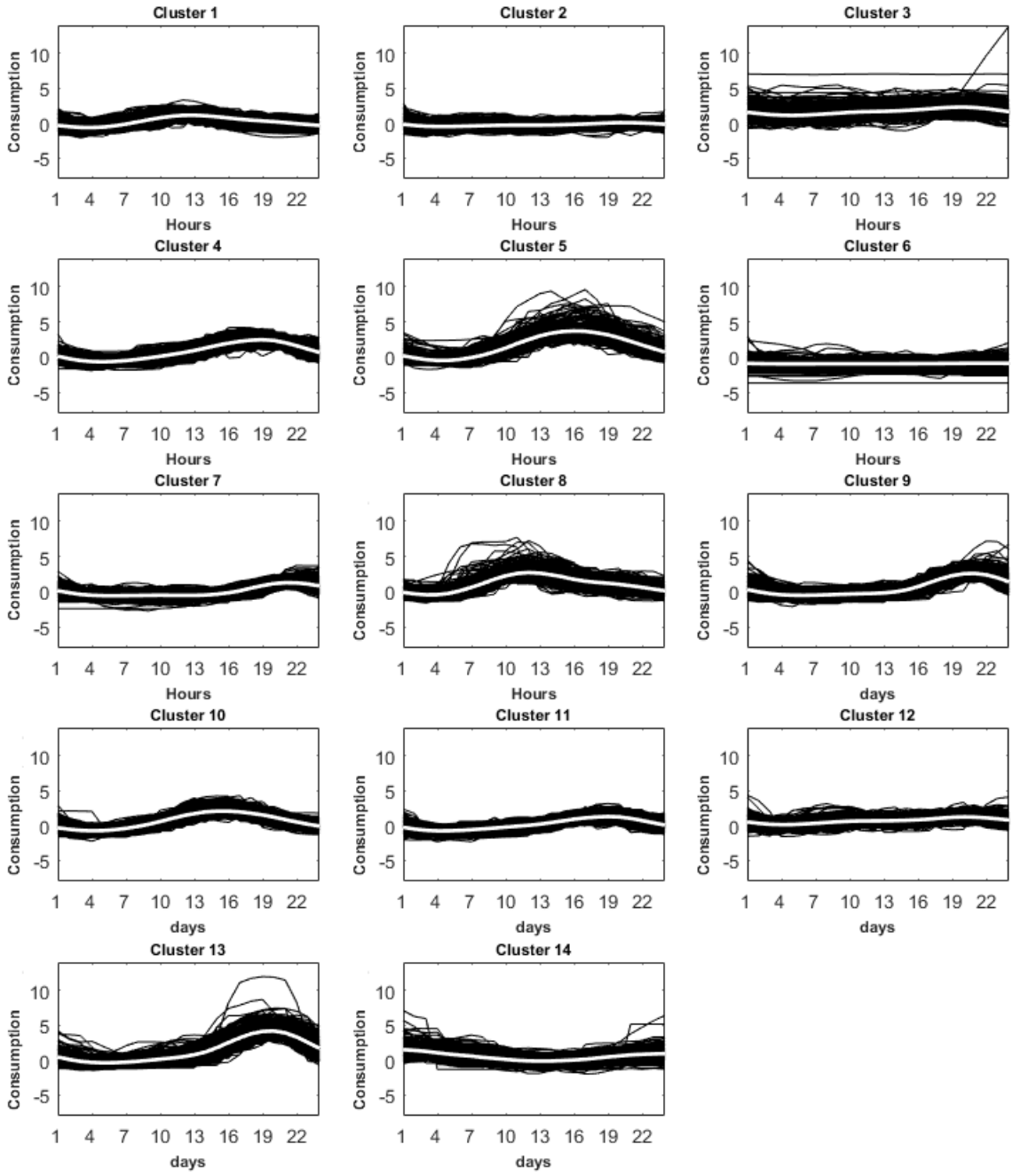


Figure 16. Case 2: 14 daily weekday profiles.

4.8.1 Similarity measure for annual profiles

The similarity of the annual load profiles based on the cluster representation in Figure 11, 12, 13 and 14 was derived by computing a two-sample Kolmogorov-Smirnov test. The Kolmogorov-Smirnov test is a non-parametric test and it confirms whether two samples come from the same distribution. The strategy is to use the Kolmogorov-Smirnov test to

detect how close the centroids of the cluster representation produced are (i.e, the white line or curves in the plots). The null hypothesis is that both profiles are sampled from the same distribution against the alternative hypothesis that they are not sampled from the same distribution. Table 7, 8, 9 and 10 show the similarity matrix formed by using the Kolmogorov-Smirnov test. The cases that seem to be the potential solution, i.e, 4, 8, 12 and 15 clusters are examined respectively. A 0 means the null hypothesis is accepted and 1 means the null hypothesis is rejected.

Table 7. CASE 1: Similarity measure for 4 annual load profiles.

Cluster	1	2	3	4
1	0	1	1	1
2	1	0	1	1
3	1	1	0	1
4	1	1	1	0

Table 8. CASE 2: Similarity measure for 8 annual load profiles.

Cluster	1	2	3	4	5	6	7	8
1	0	1	1	1	1	1	1	1
2	1	0	1	1	1	1	1	1
3	1	1	0	1	1	1	1	1
4	1	1	1	0	1	1	1	1
5	1	1	1	1	0	1	0	1
6	1	1	1	1	1	0	1	1
7	1	1	1	1	0	1	0	1
8	1	1	1	1	1	1	1	0

Table 9. CASE 3: Similarity measure for 12 annual load profiles.

Cluster	1	2	3	4	5	6	7	8	9	10	11	12
1	0	1	1	1	1	1	1	1	1	1	1	1
2	1	0	1	1	1	1	1	1	1	1	1	1
3	1	1	0	1	1	1	1	1	1	1	1	1
4	1	1	1	0	1	1	1	1	1	1	1	1
5	1	1	1	1	0	1	1	1	1	1	1	1
6	1	1	1	1	1	0	1	1	1	1	1	1
7	1	1	1	1	1	1	0	1	1	1	1	1
8	1	1	1	1	1	1	1	0	1	1	1	1
9	1	1	1	1	1	1	1	1	0	1	1	1
10	1	1	1	1	1	1	1	1	1	0	1	1
11	1	1	1	1	1	1	1	1	1	1	0	0
12	1	1	1	1	1	1	1	1	1	1	0	0

Table 10. CASE 4: Similarity measure for 15 annual load profiles.

Cluster	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1
3	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1
4	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1
5	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1
6	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1
7	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1
8	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1
9	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1
10	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1
11	1	1	1	1	1	1	1	1	1	1	0	0	1	1	1
12	1	1	1	1	1	1	1	1	1	1	0	0	1	1	1
13	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1
14	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1
15	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0

In case 1, i.e, Table 7, the centroids of the respective user profiles do not come from the same distribution according to the analysis. The value 1 in the table indicates that the null hypothesis is rejected. In case 2, i.e, Table 8 the centroids of Cluster 1, 2, 3, 4, 6 and 8 do not result from the same distribution as well. However, the scenario is different when the centroids of Cluster 7 and 5 are examined. As seen in Table 8 they are similar. In case 3 and 4 i.e, Table 9 and Table 10 respectively two cluster centroids are similar in each case i.e, Cluster 11 and 12. The remaining cluster centroids do not display any similarities according to the Kolmogorov-Smirnov test. The purpose of the Kolmogorov-Smirnov

two-sample test is to drive the appropriate number of clusters through the measure of similarities between the cluster centroids. If the distribution of two clusters centroids under consideration are identical, then the idea is that they should perhaps be merged to form one cluster. For the annual load profiles, this test does not seem to be the right way of measuring the similarities of the clusters given that as the number of clusters increase, we expect the similarity between the clusters formed to be great. This is not the case here and one possible explanation is that the lack of similarities might be as a result of the averaging at the preprocessing stage. However, as a start, the appropriate number of clusters for the annual load profiles cannot be 4 mainly because the second potential number for the appropriate number of clusters (8 clusters) produced new and unique profiles that were not seen before. The remaining possible numbers (12 clusters and 15 clusters) are big to represent the electricity consumption behaviour of the customers, the cluster representation of these users produced load profiles that are similar in shape and behaviour. Finally, the WCSS drop for the annual load profiles is not large around 8, also for the potential numbers studied, 8 is a good number and the uniqueness of the profiles created using 8 clusters gave assurance to this choice. For these reasons, 8 is chosen as the appropriate number of clusters for the annual load profiles.

4.8.2 Similarity measure for daily load profiles

The similarity measure for the daily profiles was performed using the Kolmogorov-Smirnov test. This is the same test done for the annual load profiles. Table 11 and 12 show the relationship of the cluster representation based on the cases considered as the appropriate number of clusters, i.e, 6 clusters and 14 clusters respectively. For each of these cases, the daily load profiles were generated using 6 and 14 clusters. For each case of the cluster representation, the similarity between the clusters in each case was studied using the Kolmogorov-Smirnov test.

Table 11. CASE 1: Similarity measure for 6 daily load profiles.

Cluster	1	2	3	4	5	6
1	0	1	1	0	1	1
2	1	0	1	1	1	1
3	1	1	0	1	1	1
4	0	1	1	0	0	1
5	1	1	1	0	0	1
6	1	1	1	1	1	0

Table 12. CASE 2: Similarity measure for 14 daily load profiles.

Cluster	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	0	1	1	1	1	1	0	1	0	0	0	1	1	0
2	1	0	1	1	1	1	0	1	1	1	1	1	1	1
3	1	1	0	1	1	1	1	1	1	1	1	1	1	1
4	1	1	1	0	0	1	0	0	0	0	0	0	0	0
5	1	1	1	0	0	1	1	0	0	1	1	1	0	1
6	1	1	1	1	1	0	1	1	1	1	1	1	1	1
7	0	0	1	0	1	1	0	1	0	0	0	1	1	1
8	1	1	1	0	0	1	1	0	0	0	1	1	0	1
9	0	1	1	0	0	1	0	0	0	0	0	1	0	1
10	0	1	1	0	1	1	0	0	0	0	0	0	0	0
11	0	1	1	0	1	1	0	1	0	0	0	1	1	1
12	1	1	1	0	1	1	1	1	1	0	1	0	1	0
13	1	1	1	0	0	1	1	0	0	0	1	1	0	1
14	0	1	1	0	1	1	1	1	1	0	1	0	1	0

In case 1, i.e, Table 11, the centroids of Cluster 4 and 5 are similar according to the Kolmogorov-Smirnov test. The remaining user profiles are different according to the same test. In the second case of the daily load profiles, i.e, Table 12, the similarity between these load profiles are higher. An increase in the number of profiles results in users with similar electricity consumption patterns ending up in different clusters. This explains the high similarity in Table 12. Cluster 3 and 6 are the only distinct profiles in the table, the rest of the clusters have one or more users profiles they are similar to. Due to the high similarity in Table 12 the appropriate number of clusters for the daily load profiles cannot be 14 clusters. This leaves only one choice for the appropriate number of clusters, i.e, the clusters in Table 11 . The similarity measure is intuitive in the case of the daily load profiles. Cluster 4 and 5 in Table 11 are not only similar according to the Kolmogorov-Smirnov 2 sample test but also have a similar pattern of consumption. These load profiles can be merged. Hence, the appropriate number of clusters for the daily load profiles is, therefore, 5 clusters.

4.8.3 Annual weekend load profile

The annual weekend load profiles were created using 8 clusters (the appropriate number of clusters for the annual load profiles is 8). The electricity consumption during the weekdays for the load profiles that are considered to indicate residential households is presumed to vary considerably from the weekend's consumption patterns (working class

resident spend more time at home during the weekends). For the industrial and commercial customers, the production rate is lesser during the weekends (businesses open and close production differently during the weekends). The electricity consumption pattern is also impacted by this. As a consequence, the annual weekend load profiles were obtained and compared to the annual weekday load profiles already addressed in Figure 12.

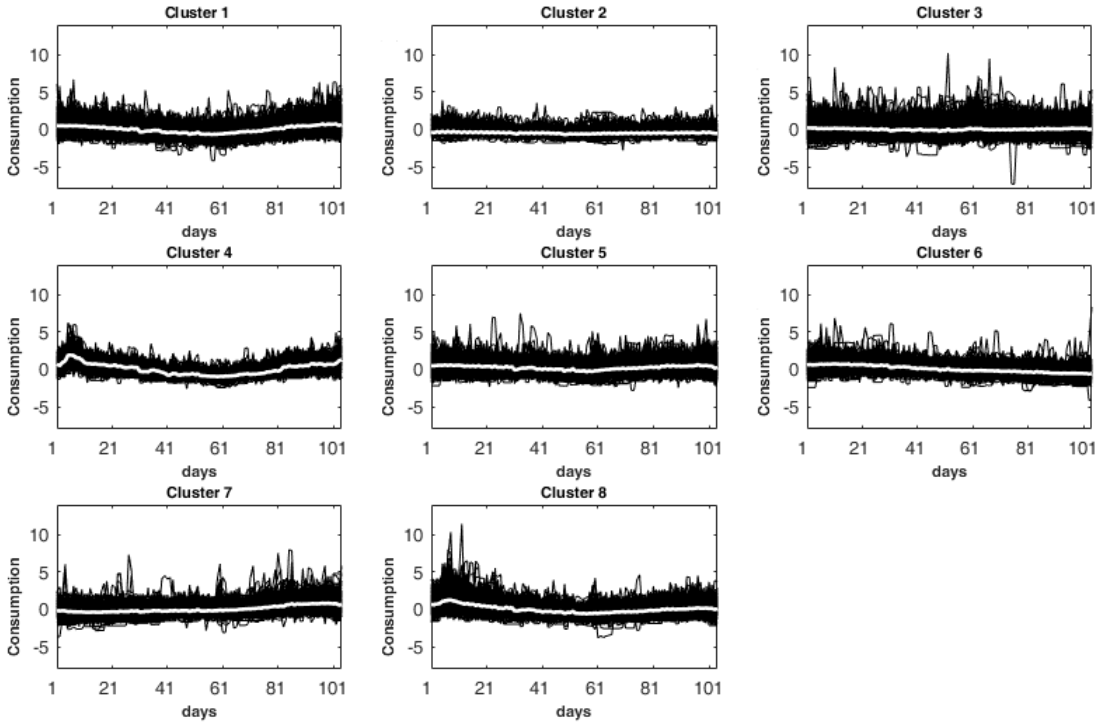


Figure 17. Annual weekend load profiles.

Ignoring the position of the clusters in the respective figures, the weekend load profiles and the weekday profiles are comparable. However, some differences can be noted. In Cluster 2 of the weekend load curves, the electricity consumption is below 5 kw whereas in the weekday load profiles many of the profiles have consumption rates above 5 kw. Also, all the users in the weekend profile have consumption below 10 kw except for few users in Cluster 3 and Cluster 8. In the weekday profiles many of the profiles have users whose power consumption rates are above 10 kw. The consumption of energy is higher during the weekday in comparison to the weekends.

4.8.4 Refining annual load profiles

As seen previously, the number 4, 8, 12 and 15 were identified as the potential value for the appropriate number of clusters in the case of the annual load profiles. The cluster representation for each of these cases have been presented already and the characteristics of the respective user profiles elaborated. In this subsection, the cluster representation of these numbers is replicated using a different method. The clusters are formed by providing the initial centroids to the K-means clustering algorithm rather than relying on the random initial centroids selection as was the case previously. To ensure that the global optimum is attained, these initial centroids were optimised using the GA. The clusters for the various cases of the annual load profiles are given in Figure 18, 19, 20, 21.

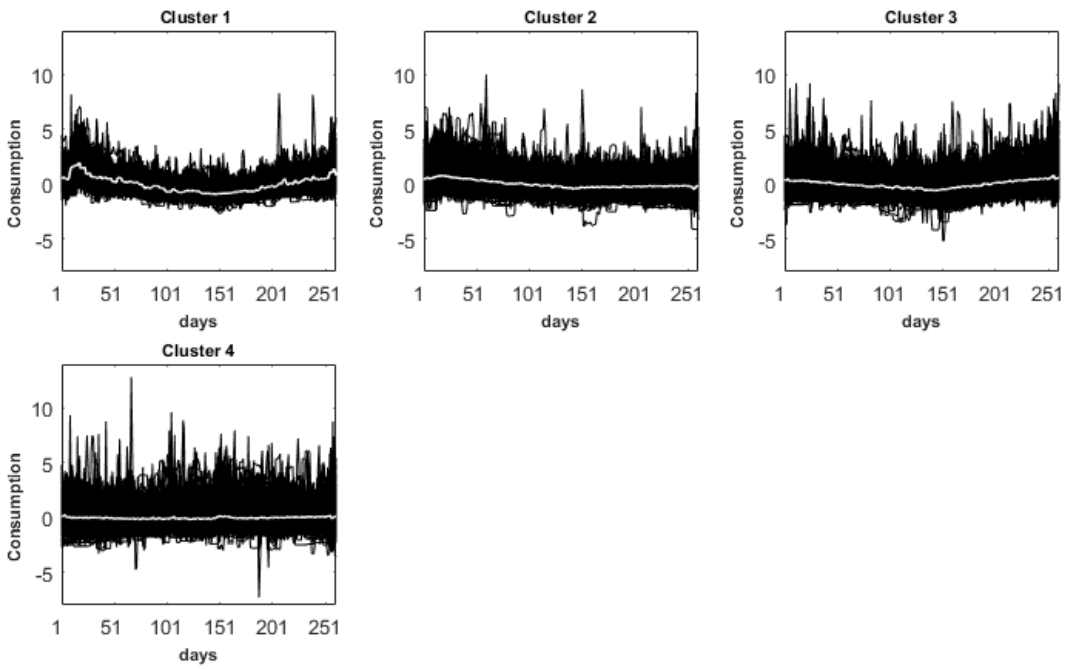


Figure 18. 4 Refined annual load profiles.

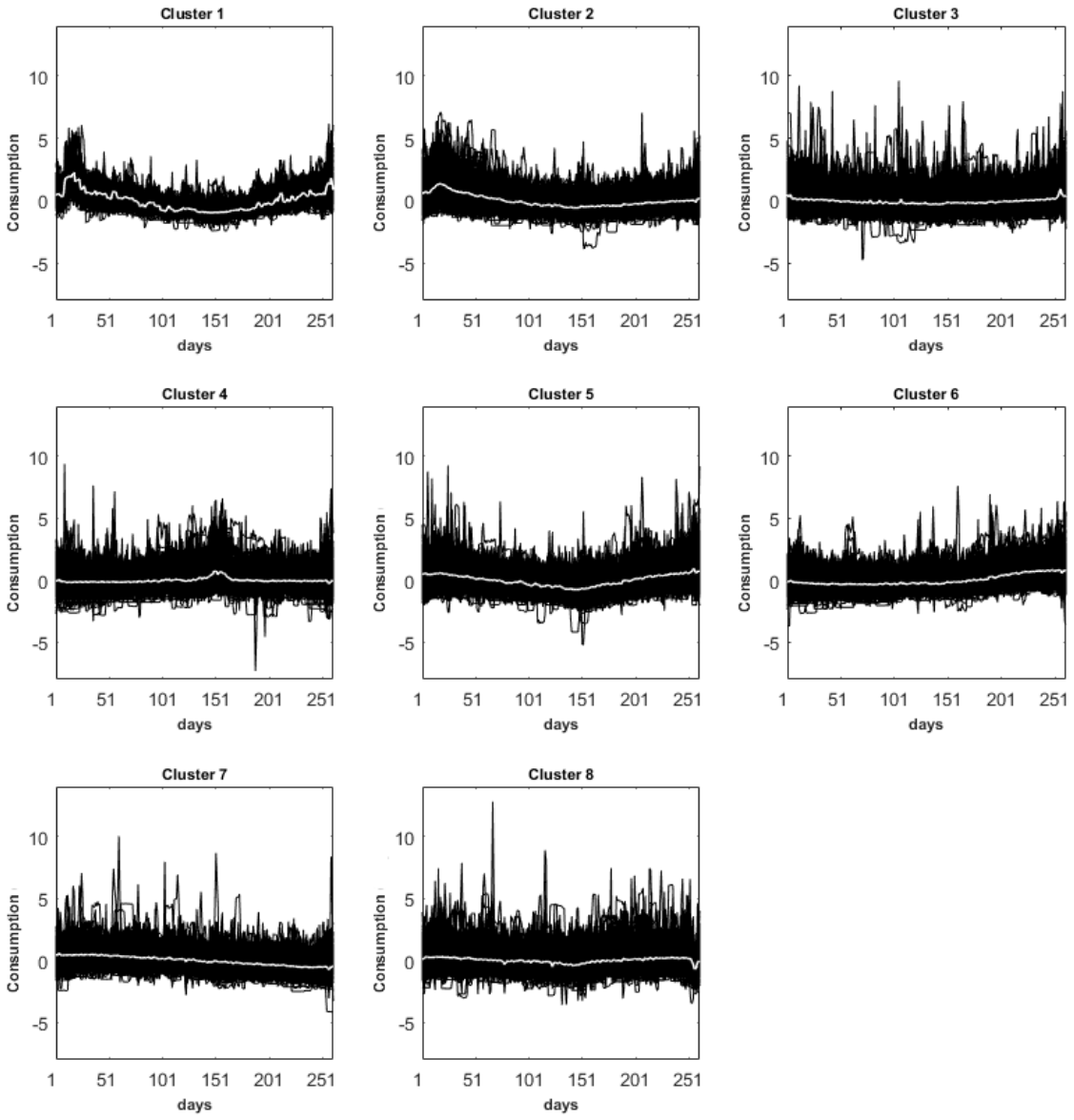


Figure 19. 8 Refined annual load profiles.

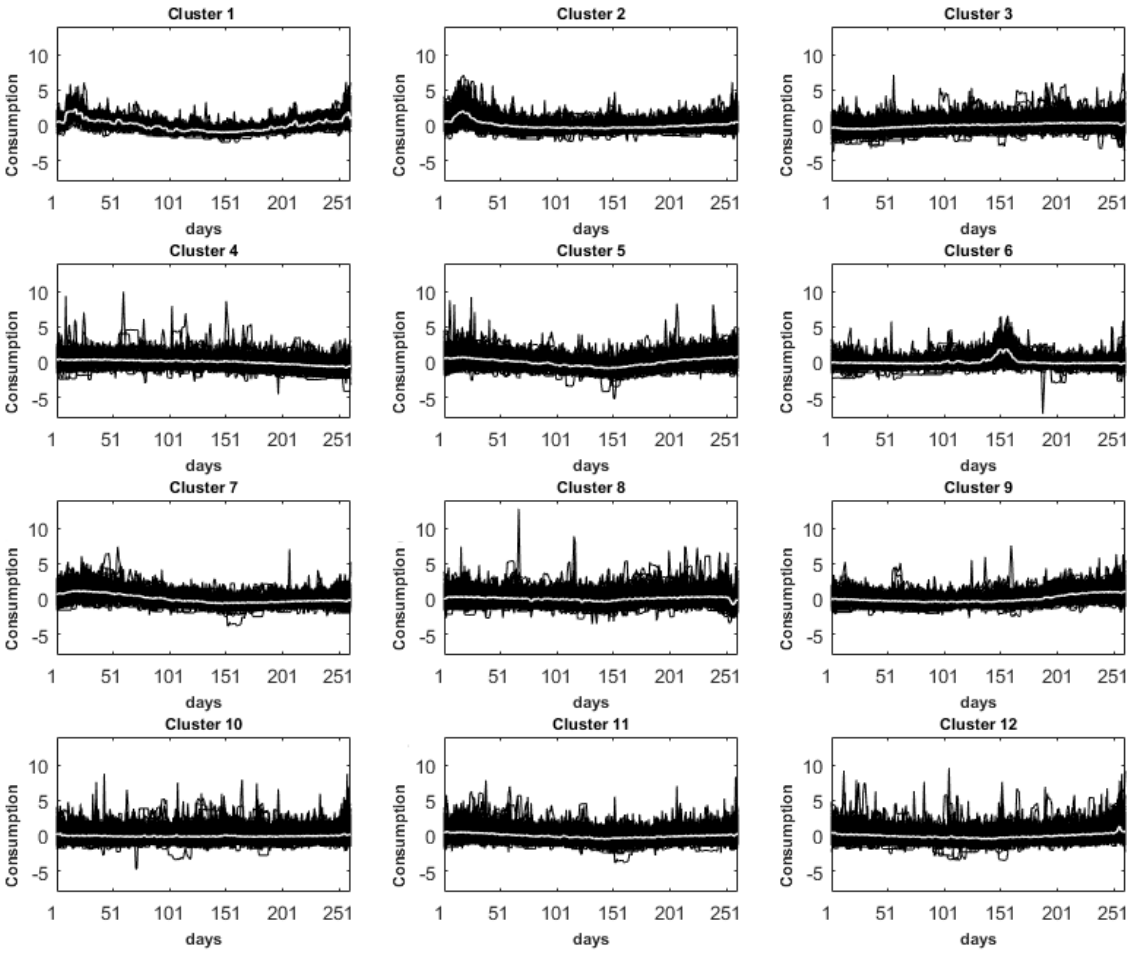


Figure 20. 12 Refined annual load profiles.

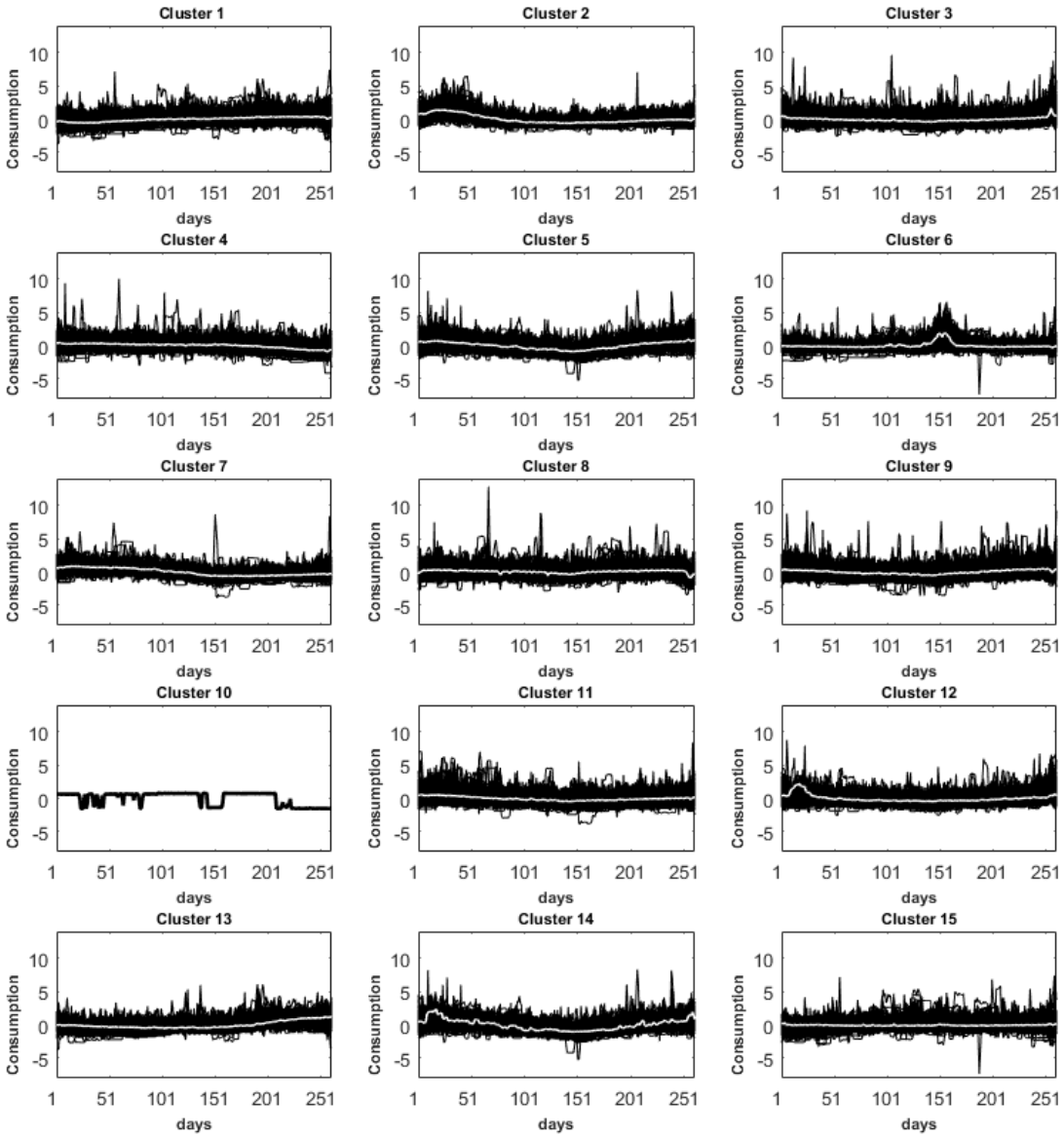


Figure 21. 15 Refined annual load profiles.

As expected the replication of the load profiles using a different method yielded similar profiles as before. In the first instance, for example, the four clusters produced are similar to the clusters found in Figure 11. The same implies to Figure 12, 13 and 14. The position of the load profiles might be different in each respective figure, but these profiles look alike. In Figure 21, Cluster 10 is empty (the cluster contains only the initial centroids value). Figure 14 has the same number of clusters as Figure 21 however, no empty clusters are produced in the former. This is one noted different.

4.8.5 Refining daily load profiles

The daily load profiles were also replicated using the same method described for the annual load profiles. In the annual load profiles replication, it has been seen that the shapes of the load profiles or pattern of electricity consumption did not change despite the method used. The same characteristics were also exhibited during the daily load profiles refining. Figure 22 and 23 contain the refine daily profiles. The load profiles were generated using the potential appropriate number of clusters discussed earlier (6 and 14 clusters).

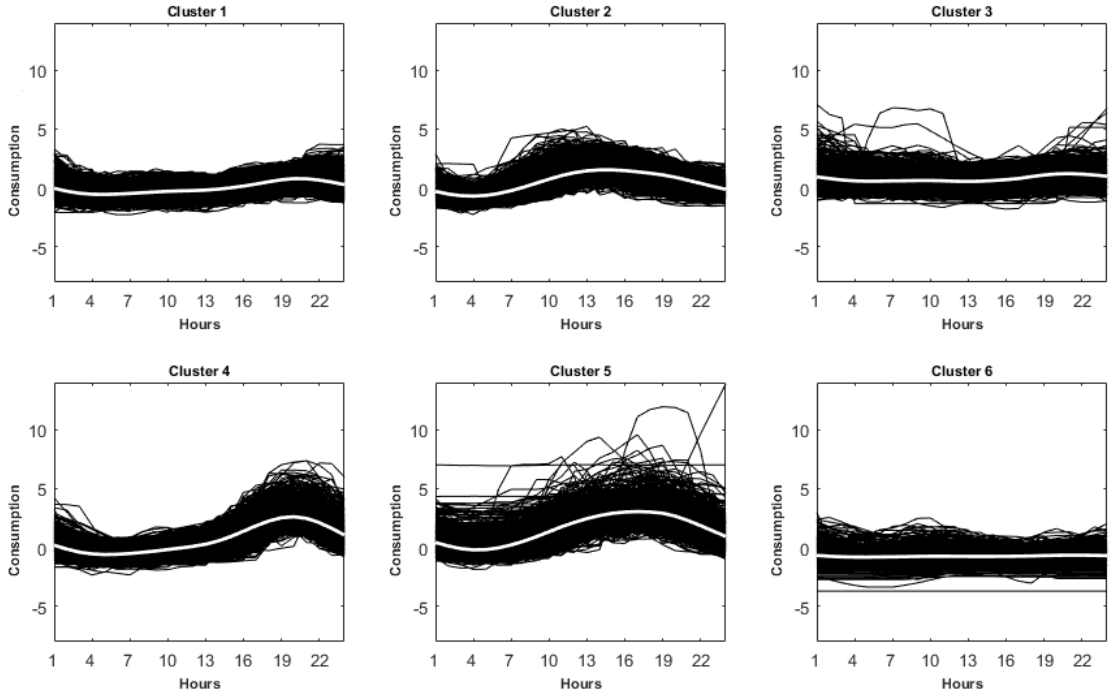


Figure 22. Case 1: 6 Refined daily load profiles.

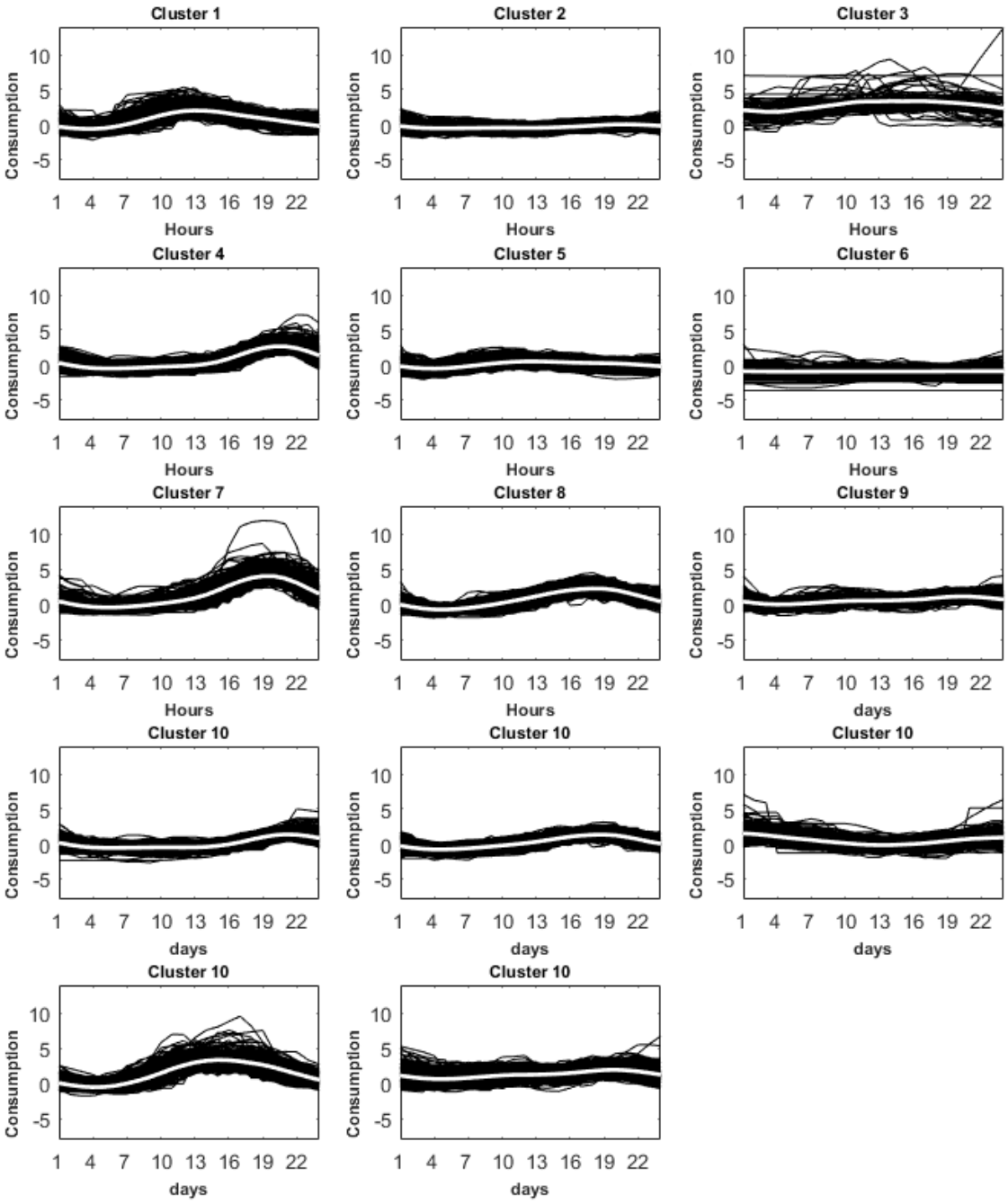


Figure 23. Case 1: 14 Refined daily load profiles.

4.9 Method comparison

The two methods were compared in order to determine if the optimization process will, in fact, improve the clustering result. Since the clusters produced via the different methods are similar visually, a way is needed to compare the performance of the methods. The validity index values (Silhouette and Davies-Bouldin) for the two methods were computed and compared. Each of the cases of the annual load profiles were compared to their corresponding refined clusters. Table 13, 14, 15 and 16 give the validity index of the two methods.

Table 13. Case 1 with 4 annual load profiles: Results of the two validity indices. Bold numbers show the best index value.

Methods	**Silhouette**	**Davies–Bouldin**
Kmeans with GA	**0.1136**	**3.0778**
Kmeans	0.1136	3.0778

Table 14. Case 2 with 8 annual load profiles: Results of the two validity indices. Bold numbers show the best index value.

Methods	**Silhouette**	**Davies–Bouldin**
Kmeans with GA	**0.0741**	**3.2759**
Kmeans	0.0741	3.2816

Table 15. Case 3 with 12 annual load profiles: Results of the two validity indices. Bold numbers show the best index value.

Methods	**Silhouette**	**Davies–Bouldin**
Kmeans with GA	**0.0590**	**3.3956**
Kmeans	0.0580	3.4046

Table 16. Case 4 with 15 annual load profiles: Results of the two validity indices. Bold numbers show the best number index value.

Methods	**Silhouette**	**Davies–Bouldin**
Kmeans with GA	**0.0568**	**3.2626**
Kmeans	0.0508	3.6130

From Table 13, the validity indices for the two methods are the same. This means refining the clusters did not improve or decrease the quality of the clustering results in that case. As the number of clusters increased it is noticed that refining the clusters slightly improved the quality of the clustering result. This can be observed in Table 14, 15 and 16. The two validity indices test did indeed support our assumption that the optimisation of the initial centroids values using GA will improve the quality of the load clustering results.

5 DISCUSSION

During the experiment, an emphasis was placed on clustering the load data into their respective groups. These profiles have potential in helping the energy company in the identification of unknown user electricity profiles. The load profiles can also be very useful in studying the behaviour of known profiles in the case that the consumption pattern changes from the known behaviours. The load profiles derived from this study can be useful in modelling the new pattern of behaviour by associating the pattern to already known behaviours formed during the clustering process. And finally, the clusters can serve as initial representatives for a classification algorithm that requires the availability of a training data set. To this regard, the accuracy of the clusters is important.

The accuracy of the clusters depends largely on selecting the appropriate number of clusters. Similar studies suggested different numbers for the appropriate number of clusters. The number of clusters ranges from 10, 15 to 20 for some studies and as low as 3 for the others. Considering these studies are not conducted in Finland (the electricity consumption can be affected by the geographical location), a method for selecting the right number of cluster for this case study was needed. Selecting the number of clusters presented a challenge for this study. Similar work conducted in Finland dates back to the $1980s$ (this study has not been able to find a recent work done in Finland were the number of clusters has been suggested) and the consumption behaviour is believed to have changed ever since. Examining the clusters formed using the potential number of clusters for the different cases considered and the WCSS to some extent as well as with the Kolmogorov-Smirnov two sample test to study the similarity of the clusters, visually the selected clusters exhibit unique characteristics. This has given confidence to the choice of the appropriate number of clusters considering the validity indices initially suggested a very low number as the appropriate number of clusters.

The main motivation of using the GA was for improving the quality of the load profiles formed through the K-means clustering. This is very important in ensuring that a user is clustered in the right group. The energy company cannot create plans for every single user, clustering the users into groups will mean energy plans are created for similar users. These plans cannot be sensitive. If users are assigned to the wrong cluster then the plans cannot be trusted. Also, in creating plans for predicting the electricity usage of a customer, the user has to be in the right group for this plan to work. This is the benefits GA ensures. The annual cluster quality are shown to be improved in certain cases by initial centroid optimization using GA.

From the different profiles formed after the clustering process, we see differences in the characteristics exhibited by each respective profile. In the daily profile, similarities exist in the consumption behaviour shown in the early hours of the morning, the demand for electricity during these hours are shown to be very low. This behaviour is shown by all profiles believed to indicate residential homes. Some daily load profiles showed some evidence of a peak in the consumption rates during a certain time of the day. These peaks occurred at fairly different hours showing how the demand for electricity differs in each profile. Also, most of the daily profiles formed exhibited a decrease in the need for electricity before midnight. Most of the profiles showed this common behaviour. This shows that the demand for electricity by the customers in Southern Finland is limited at night and the early mornings compared to the daytime.

The annual profiles also showed some unique characteristics. In the winter season, for instance, residential homes with no district heating showed a higher rate of consumption in colder days in comparison to homes where district heating is believed to be present. In June, most of the profiles showed a decrease in energy consumption. Few annual profiles showed no difference in the consumer behaviour across the entire year. This was believed to indicate the industrial or commercial establishments. In a supermarket, for example, the rate of electricity needed to preserve the food commodities is always the same despite the season. This explains the consistent behaviour of these profiles. The evidence of the colder and warmer seasons are clearly exhibited by some of the load profiles formed.

The daily profiles showed similar behaviour across the week. The daily peaks are seen to occur approximately at the same time during the day across the entire week. Based on this knowledge, the prediction of a user's electricity usage with similar consumption behaviour as the load profiles already created can be easily achieved. This fact is clearly shown when the weekly profiles are considered.

6 CONCLUSION

In this master's thesis, different methods for electricity load profiling was studied. A common method to cluster a large load data into groups with similar characteristic was chosen. A popular data mining algorithm called the K-means clustering algorithm was used in achieving this objective. The appropriate number of clusters was verified through the use of the Davies-Bouldin validity index and the WCSS to some extent.The load data was successful group together in a cluster, in which each cluster consisted of electricity users with similar consumption behaviour. Considering some of the downsides this algorithm is known to exhibit an evolutionary genetic algorithm was proposed to complement the K-means clustering. The results show that the GA has, in fact, slightly increase the accuracy of the clustering results in certain cases.

REFERENCES

[1] Yi Sun, Wei Gu, J. Lu, and Zenghui Yang. Fuzzy clustering algorithm-based classification of daily electrical load patterns. In *2015 12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, pages 50–54. IEEE, Aug 2015.

[2] A. Mutanen, P. Järventausta, M. Kärenlampi, and P. Juuti. Improving distribution network analysis with new AMR-based load profiles. In *22nd International Conference and Exhibition on Electricity Distribution (CIRED 2013)*, pages 1–4, June 2013.

[3] Alexander Lavin and Diego Klabjan. Clustering time-series energy data from smart meters. *Energy efficiency*, 8(4):681–689, 2015.

[4] P. McDaniel and S. McLaughlin. Security and privacy challenges in the smart grid. *IEEE Security Privacy*, 7(3):75–77, May 2009.

[5] G. Chicco, R. Napoli, and F. Piglione. Load pattern clustering for short-term load forecasting of anomalous days. In *2001 IEEE Porto Power Tech Proceedings (Cat. No.01EX502)*, volume 2, page 6 pp., 2001.

[6] G. Chicco, R. Napoli, P. Postolache, M. Scutariu, and C. Toader. Customer characterization options for improving the tariff offer. *IEEE Transactions on Power Systems*, 18(1):381–387, Feb 2003.

[7] 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*, 7(1):136–144, Jan 2016.

[8] Electricity Association. Load profiles and their use in electricity settlement, 1997.

[9] Zuhaina Hj Zakaria and K. L. Lo. Load profiling in the new electricity market. In *Student Conference on Research and Development*, pages 278–281, 2002.

[10] A. H. Nizar, Z. Y. Dong, and J. H. Zhao. Load profiling and data mining techniques in electricity deregulated market. In *2006 IEEE Power Engineering Society General Meeting*, pages 7 pp.–, 2006.

[11] S. V. Verdu, M. O. Garcia, F. J. G. Franco, N. Encinas, A. G. Marin, A. Molina, and E. G. Lazaro. Characterization and identification of electrical customers through the use of self-organizing maps and daily load parameters. In *IEEE PES Power Systems Conference and Exposition, 2004*, pages 899–906, Oct 2004.

[12] Gianfranco Chicco, Roberto Napoli, Federico Piglione, Petru Postolache, Mircea Scutariu, and Cornel Toader. A review of concepts and techniques for emergent customer categorisation. In *TELMARK Discussion Forum European Electricity Markets, London*. Citeseer, 2002.

[13] Pavel Berkhin. A survey of clustering data mining techniques. In *Grouping multidimensional data*, pages 25–71. Springer, 2006.

[14] P Praveen and B Rama. Improving efficiency and effectiveness of hierarchical clustering. *International Journal*, 9(2), 2018.

[15] Wang Peizhuang. Pattern recognition with fuzzy objective function algorithms (james c. bezdek). *SIAM Review*, 25(3):442, 1983.

[16] G. Chicco, O. M. Ionel, and R. Porumb. Electrical load pattern grouping based on centroid model with ant colony clustering. *IEEE Transactions on Power Systems*, 28(2):1706–1715, May 2013.

[17] 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*, 19(2):1232–1239, May 2004.

[18] Ran Li, Furong Li, and Nathan D Smith. Multi-resolution load profile clustering for smart metering data. *IEEE Transactions on Power Systems*, 31(6):4473–4482, 2016.

[19] C. Mihai, D. Ilea, and P. M. Mircea. Use of load profile curves for the energy market. In *2016 International Conference on Development and Application Systems (DAS)*, pages 63–70. IEEE, May 2016.

[20] D. Colley, N. Mahmoudi, D. Eghbal, and T. K. Saha. Queensland load profiling by using clustering techniques. In *2014 Australasian Universities Power Engineering Conference (AUPEC)*, pages 1–6, Sept 2014.

[21] Y. Wang, Q. Chen, C. Kang, and Q. Xia. Clustering of electricity consumption behavior dynamics toward big data applications. *IEEE Transactions on Smart Grid*, 7(5):2437–2447, Sept 2016.

[22] Fintan McLoughlin, Aidan Duffy, and Michael Conlon. A clustering approach to domestic electricity load profile characterisation using smart metering data. *Applied energy*, 141:190–199, 2015.

[23] Teemu Räsänen, Dimitrios Voukantsis, Harri Niska, Kostas Karatzas, and Mikko Kolehmainen. Data-based method for creating electricity use load profiles using

large amount of customer-specific hourly measured electricity use data. *Applied Energy*, 87(11):3538–3545, 2010.

[24] Teemu Räsänen, Juhani Ruuskanen, and Mikko Kolehmainen. Reducing energy consumption by using self-organizing maps to create more personalized electricity use information. *Applied Energy*, 85(9):830–840, 2008.

[25] James MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA, 1967.

[26] Ioannis P Panapakidis and Georgios C Christoforidis. Implementation of modified versions of the k-means algorithm in power load curves profiling. *Sustainable Cities and Society*, 35:83–93, 2017.

[27] Oliver Kramer. *Genetic algorithm essentials*, volume 679. Springer, 2017.

[28] Hanan Samet. *Foundations of multidimensional and metric data structures*. Morgan Kaufmann, 2006.

[29] Eamonn J Keogh and Michael J Pazzani. A simple dimensionality reduction technique for fast similarity search in large time series databases. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 122–133. Springer, 2000.

[30] Yi Wang, Qixin Chen, Chongqing Kang, Mingming Zhang, Ke Wang, and Yun Zhao. Load profiling and its application to demand response: A review. *Tsinghua Science and Technology*, 20(2):117–129, 2015.

[31] Slobodan Petrovic. A comparison between the silhouette index and the davies-bouldin index in labelling IDS clusters. In *Proceedings of the 11th Nordic Workshop of Secure IT Systems*, pages 53–64, 2006.

[32] Bernard Desgraupes. Clustering indices. *University of Paris Ouest-Lab Modal'X*, 1:34, 2013.