A pattern recognition methodology for analyzing residential customers load data and targeting demand response applications

Amin Rajabi[1,*], Mohsen Eskandari[1], Mojtaba Jabbari[1], Sahand Ghavidel[1], Li Li[1], Jiangfeng Zhang[1], Pierluigi Siano[2]

1=Faculty of Engineering and Information Technology, University of Technology Sydney, PO Box 123, Broadway, NSW 2007, Australia

2= Department of Industrial Engineering, University of Salerno, Via Giovanni Paolo II, 132, Fisciano (SA) 84084, Italy.

Abstract

The availability of smart meter data allows defining innovative applications such as demand response (DR) programs for households. However, the dimensionality of data imposes challenges for the data mining of load patterns. In addition, the inherent variability of residential consumption patterns is a major problem for deciding on the characteristic consumption patterns and implementing proper DR settlements. In this regard, this paper utilizes a data size reduction and clustering methodology to analyze residential consumption behavior. Firstly, the distinctive time periods of household activity during the day are identified. Then, using these time periods, a modified symbolic aggregate approximation (SAX) technique is utilized to transform the load patterns into symbolic representations. In the next step, by applying a clustering method, the major consumption patterns are extracted and analyzed. Finally, the customers are ranked based on their stability over time. The proposed approach is applied on a large dataset of residential customers' smart meter data and can achieve three main goals: 1) it reduces the dimensionality of data by utilizing the data size reduction, 2) it alleviates the problems associated with the clustering of residential customers, 3) its results are in accordance with the needs of systems operators or demand response aggregators and can be used for demand response targeting. The paper also provides a thorough analysis of different aspects of residential electricity consumption and various approaches to the clustering of households which can inform industry and research activity to optimize smart meter operational use.

Keywords

Smart meters, Load patterns, Demand response, Clustering algorithms, Symbolic aggregate approximation (SAX).

[*] Corresponding author: amin.rajabi@student.uts.edu.au

1. Introduction

Smart meters have been vastly deployed for residential premises in recent years. These meters are equipped with measurement and communication capabilities that enable them to record the fine-grained energy consumption of customers and provide added information to the utility company. It is projected that the total number of installed smart meters will reach 780 million in 2020 [1]. This increasing trend has fundamentally changed the interaction between utilities and electricity customers and opened up possibilities for companies to offer new services such as customized tariff structures and demand response (DR) programs to the users.

Smart meters can record the consumption data at different time resolutions, obtaining the daily load curve (DLC) of each customer. Therefore, there is a DLC for the customer *i* and the day *j*. The availability of this huge amount of DLCs offers great opportunities to study the consumption behavior of residential customers. For example, data mining techniques can be applied on the consumption data to characterise the relevant features of consumption behavior. Clustering is one of the most popular techniques in data science. The aim of clustering is to identify similar groups in a dataset. Objects belonging to one cluster are more similar to each other than to those in other clusters. By applying clustering techniques on DLCs of one customer, similar DLCs are grouped into one cluster and the major consumption behaviors of the customer are identified. On the other hand, clustering can also help to segregate a customer set into several clusters where the customers belonging to each group exhibit similar consumption patterns.

Clustering can assist in establishing residential DR programs since it divides different customers or different load patterns into distinctive classes. This process can be used to target suitable groups by appropriate DR programs. For example, those customers with load peaks happening at the same time as system peak are good prospects to be targeted by higher electricity prices or incentives to shift their loads from peak hours to off-peak hours. On the other hand, clustering can be used in more sophisticated ways to enhance the effectiveness of DR plans. However, in the literature, the use of clustering methods as a preliminary step in DR establishment has not been explored in detail. Only a limited number of studies have examined the application of clustering for facilitating and improving residential DR.

In this paper, a data size reduction and clustering approach is utilized to analyze the DLCs of residential customers. The study in this work can be seen as a new approach for filling a gap in residential customer clustering and also a new method along with the current proposals for demand response targeting. Therefore, the aim of the paper is twofold: firstly, to develop and apply a symbolic aggregate approximation (SAX) technique, as a proper data size reduction method, on a large number of DLCs of the residential customers and analyze their consumption patterns; and secondly, to apply the results for residential DR programs . In spite of its efficacy, the application of SAX for the residential customers and a large number of DLCs has not been explored in detail. To bridge this gap, this study offers several contributions as follows:

- It investigates the application of a modified SAX technique on a large dataset comprising hundreds of thousands of DLCs of residential dwellings. Use of SAX is appropriate since the residential DLCs usually display high variability from one day to another day. Therefore, instead of using the DLCs, the relevant SAX representations can be used which brings in more meaningful outcomes.
- To apply the SAX, the main time periods of household activity during the day should be identified. Therefore, an analysis of consumption data is carried out to identify the critical time periods during the day which are used in the SAX to partition the time axis.

- Using a clustering approach, the SAX representations of data are assigned to different clusters. In this stage, two modifications are applied to distance calculation to improve the clustering results. In addition, the effects of the parameters of the SAX technique and cluster numbers are studied by appropriate measures. The obtained results are analyzed which give promising insights about households' consumption patterns.
- The results of the clustering stage are utilized to help in procuring DR from the residential households. The use of SAX is helpful in this stage too. It is in accordance with the needs of retailers or DR aggregators which usually require DR or load changes in specific time periods.

The remaining parts of the paper are organized as follows. In Section 2, the extant literature on clustering of load data and cluster-based DR is briefly summarized and the difficulties in clustering of residential customers due to the high variability in consumption patterns are described. Based on this analysis, stages of the suggested method are explained. The theoretical concepts of the methodology including the data size reduction method, clustering algorithm, and entropy analysis are introduced in Section 3. In Section 4, the data set, which is used in this study, is evaluated to extract the main time periods which are used in the SAX method. The case studies and the results are presented in Section 5. Finally, conclusions are presented in Section 6.

2. Literature review and problem statement

2.1. Literature review

2.1.1. Clustering of households

Before the widespread rollout of smart meters, the clustering of residential electricity customers was mostly performed using the certain household characteristics. Therefore, instead of using the real load data, various dwelling and household attributes were used to categorize different classes of customers. Variables such as building characteristics, socio-economic factors, habits of consumption usage, and attitudes toward energy use were among the important factors for customer clustering. These data were usually obtained through in-home surveys or available social datasets. With the introduction of smart meters, research on household clustering has changed its focus from attribute-oriented to consumption-pattern-oriented approach [2]. The availability of high resolution consumption data allows clustering of households in more accurate and innovative ways.

The main techniques used for electricity customer clustering include: K-means [3], [4] and its variations such as fuzzy K-means [5], [6], hierarchical algorithms [3], [7], self-organizing maps (SOM) [8], [9], and probabilistic and generative models [10], [11]. Other algorithms such as follow-the-leader [3], density-based spatial clustering of applications with noise (DBSCAN) [12], [13], and K-shapes [14] are examined in some publications. Since the consumption values are continuously recorded over time, dynamic [15], [16] and online [17], [18] clusterings of time series data are also considered in a few studies.

Analyzing the massive sets of DLCs from tens of thousands of smart meters could be difficult for electrical utilities. Applying clustering techniques on these data, especially when the number of customers is very big or the time period of study is long, would be a challenging task. One common solution is to assign a representative load pattern (RLP) to each customer. RLPs are usually calculated by averaging the DLCs over a period of time, for example, a season. Therefore, instead of dealing with many DLCs, every customer is represented by just one RLP. Clustering of customers can then be achieved using these RLPs. The shortcomings of using RLPs are illustrated in Section 2.2.

Another solution is to transform the DLCs of customers into new variables before applying the clustering. This preliminary stage can be performed using three main approaches: feature extraction, feature definition, and data size reduction techniques. In this way, for each customer, a limited set of variables are calculated using the original DLCs and then, clustering techniques are applied on these variables. The main benefits of such transformation can be the management of data deluge, improvement of computation time, and in some cases superior and more interpretable results.

Feature extraction and feature definition approaches construct new features from the consumption data. Techniques such as discrete Fourier transform (DFT) [19], discrete wavelet transform (DWT) [20], and frequency domain analysis [21] have been employed to extract features from the load data. In the feature definition approach, the load patterns of customers are represented by a set of features which are defined by the user [22], [11].

The third approach involves the use of data size reduction techniques. In the power system literature, principal component analysis (PCA) is vastly utilized to convert the load data to a few components [23-25]. SAX is a well-known data size reduction technique in the data mining literature [26]. It is a suitable technique for producing a representation of original time series data; however, its use for residential load data is not studied.

Table 1 reports some of the studies on clustering of electricity customers. They have various objectives and employ different clustering methods.

Table 1 Literature on clustering of consumption data using different methods and for different purposes

Ref.	Dataset	Clustering method	Descriptions	Results
[27]	3941 residential customers	K-means, K-medoids, SOM	A multinomial logistic regression is used to link the household characteristics to the customer segmentation	Understanding the effect of different variables on cluster membership
[13]	31 residential customers	DBSCAN	A mixed integer non-linear programming is used for solving the optimization problem	Finding customized retail prices for different classes of customers
[10]	Ameren Illinois database (the number of customers is not specified)	K-means, Hierarchical, Gaussian mixture model	Use of RLPs	Obtaining optimal time of use (TOU) structures
[28]	103 residential customers	K-means	Use of RLPs	Finding the correlation between the household characteristics and load patterns
[29]	3440 residential customers	K-means	Use of RLPs Attributes are selected from the survey data for classification	Classification of new electricity customers
[30]	Residential customers: 3994 Other customers: 1341	Kernel spectral clustering	*Feature extraction:* A methodology based on wavelet feature extraction is utilized.	Improving the load forecasting
[11]	3622 residential customers	Gaussian mixture model	*Feature definition:* Seven features are defined and calculated based on the load data	To understand the peak demand and major sources of variability in customers' behavior
[31]	1824 connection points/ 18,098 customers (each connection point consists of one or more customers)	K-means, Gaussian mixture model	*Data size reduction:* PCA for understanding and visualizing the consumption data A regression analysis to find the most important explanatory factors for the load modeling	Finding the important explanatory factors which affect the electricity consumption
[32]	234 non-residential customers	Hierarchical	*Data size reduction:* Use of SAX for data size reduction Use of RLPs	Customer segmentation

2.1.2. Residential demand response targeting using the clustering

DR plans were traditionally focused on larger customers such as industrial or commercial users. Due to the widespread deployment of smart meters, there is a potential to apply similar DR programs to residential households. The DR programs are generally divided into two main categories [33]: incentive (event)-based DR and price (time)-based DR. The incentive-based programs provide users with incentives or payments during periods of system stress such as peak load, while the price-based programs offer customers time-varying electricity prices [34]. The main functions of these programs can be summarized as peak shaving that aims at reducing the peak, valley filling which corresponds to more energy use during off-peak hours, and load shifting which is a combination of both approaches and aims at shifting the electricity consumption from the peak periods to off-peak hours.

Publications on the use of clustering for DR management are limited. Table 2 reports the characteristics and findings of some of these studies.

Table 2 Literature on clustering-based DR for residential customers

Ref.	Dataset	Clustering method	Descriptions	Results
[35]	8337 dwellings	K-means	Dividing the customers into six groups based on the time of their corresponding peaks	Finding the suitable customers for reduction of system peak
[36]	Load data of 1693 customers Surveys for 500 customers	An expectation-maximization model for clustering the load data K-means for segmentation of users based on household attitudes toward DR	Survey features are used for social segmentation Measurements at appliance level (only 58 households) for up-scale of the data	Finding the potentials of wet appliances for DR
[37]	218,000 customers	Two-stage clustering using the adaptive K-means and hierarchical clusterings	The amount of usage and the variability of usage pattern are analyzed	Multi-dimensional segmentation of customers based on their quantity and variability Recommendations for energy efficiency programs and time of use shifts for different households
[38]	The consumption data of a university campus	K-means	Temperature, time of use, and load data are used as three main features for clustering	The eligible points for DR are identified in the network
[39]	1100 dwellings Socio-economic data of 950 households	Spectral clustering and K-medoids	A hidden Markov model (HMM) is utilized to detect the occupancy states	Understanding consumption dynamics of customers for DR management
[23]	1800 customers	SOM	PCA is used for data size reduction	Comparing control (400 customers) and trial (1400 customers) groups to find out the effect of DR plans on consumption

In one attempt [35], the aim is to reduce the electrical network peak by applying suitable DR programs to the customers, which is achieved by clustering the residential users into six classes. The cluster which has the most similar demand curves to the system peak demand is selected as the most viable option for DR. Ref. [36] reports the results of a project for finding the potential of wet appliances (tumble dryers, washing machines, and dish washers) for DR. A model-based clustering approach is used for the clustering of load patterns. In addition, the customer’s willingness to participate in DR is determined using the clustering of survey attributes and the potential DR is estimated using the combination of both results. Kwac *et al.* [37] analyze a large data set and use a two-step clustering methodology to build a load dictionary which contains the representative load shapes. Different aspects of customer consumption including consumption magnitude and variability are discussed which can be used for the establishment of DR programs. In [38], a hierarchical structure is proposed to alleviate the negative impacts when multiple DR plans act simultaneously on the same network. A clustering mechanism which takes into account the consumption value, time of use, and temperature is used to detect the points which are eligible for DR program application. Ref. [39] divides electricity consumption into two parts: occupancy-related and weather-related consumption. The authors suggest that when weather effects are accounted for, household consumption is solely based on the occupancy. Here, occupancy refers to socio-demographic factors and lifestyle and it is characterized by magnitude, duration, and variability of consumption.

In spite of the vast number of publications, there are still many challenges of residential DR which are not addressed carefully [34], [40], [41]. Considerations about the customer behavior, usage pattern, and willingness to participate in DR programs are among the main issues. In this respect, we will employ a method to overcome one of the problems of residential DR implementation.

2.2. Problem statement

Once the clustering of residential energy data is concerned, two problems should be addressed carefully. The first one arises from the dimensionality of the data which was highlighted in Section 2.1.1. The other is associated with the high variability of residential DLCs. The DLCs of dwellings can change significantly on a daily basis. It means that a residential user does not exactly follow the same consumption pattern from one day to another day and does not repeatedly use a specific appliance or electric equipment at the same time every day. Theoretically, a DLC can exhibit an infinite number of shapes. This is a serious obstacle in the clustering of households into distinctive groups.

Fig. 1 displays the typical energy curve of a household during a week. Fig. 2 shows the variation of the DLCs of this dwelling for all the days in different seasons.

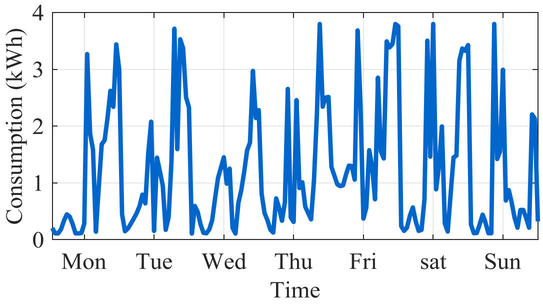

Fig. 1. Variability of the consumption of a household over a week

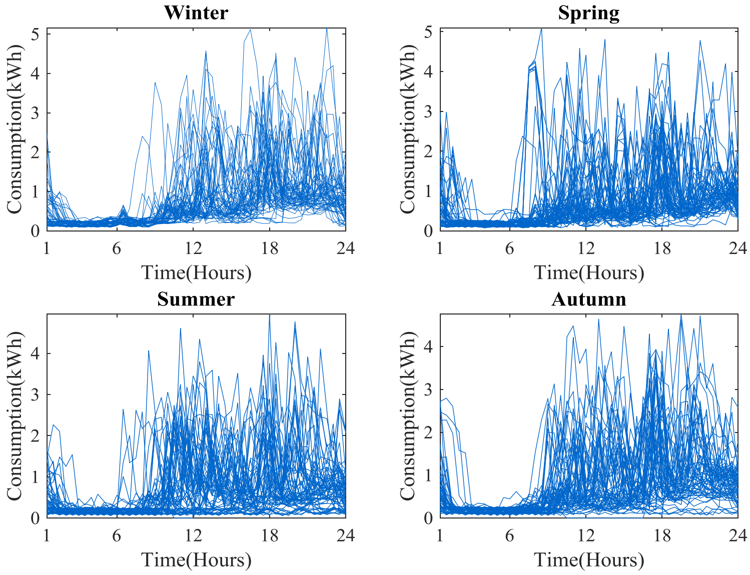

Fig. 2. Effect of seasonality on the consumption of a household

As these figures demonstrate, the consumption patterns of a residential customer can vary on a daily basis and can be largely affected by seasonal changes. This imposes challenges to the clustering of customers into a certain number of classes since the DLCs of a customer might be assigned to many different clusters. Therefore, if clustering is carried out using algorithms such as K-means and the distance measures like Euclidean distance, each customer can be assigned to many different clusters since its DLCs vary widely among different days.

As mentioned earlier, two methods, including the definition of RLPs, are followed in the literature to deal with this problem. Use of RLPs can reduce the volume of data and makes the clustering achievable. However, the RLPs cannot reflect the daily and intraday variations of energy consumption. For example, as shown in Fig. 3, two customers with completely different daily load patterns can still have similar RLPs (shown by the black lines). According to this figure, the first customer has a very unpredictable consumption behavior during the week, while the other follows a very regular pattern. It means that, regardless of the consumption value, the low and high consumption periods of the second customer do not differ considerably from one day to another. This is especially critical for DR applications when understanding the stability of user's consumption habits over time is of high importance.

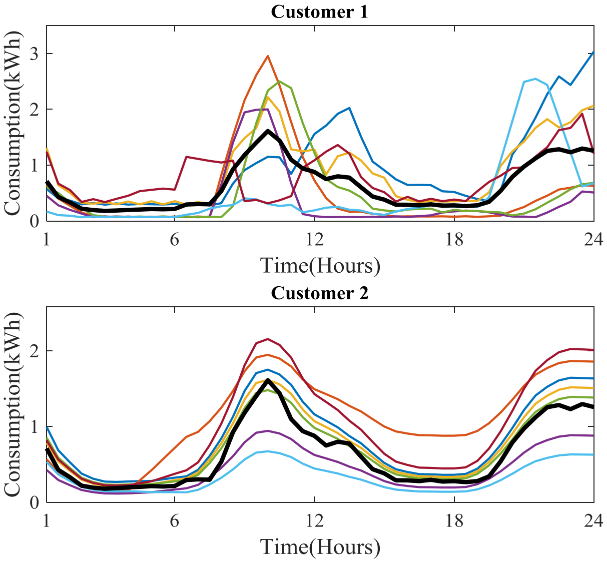

Fig. 3. Consumption behavior of two customers

Use of SAX technique can help in settling the aforementioned problems. It has two suitable features which make it a perfect tool for analyzing residential load patterns. First, it considers several important time periods during the day instead of focusing on the exact time of consumption. Secondly, instead of dealing with the infinite number of consumption values, it transforms the load data into a series of symbols. Using these two features, each load pattern can be represented by a limited number of symbols. This allows limiting the number of many DLC varieties while retaining the important information of original load data. SAX representations can then be clustered using proper clustering techniques. Furthermore, applying the SAX can facilitate DR applications for example, by enabling the segregation of stable customers from variable ones.

2.3. Stages of the method

Based on the aforementioned analysis, the proposed methodology in this paper which includes four main steps is shown in Fig. 4.

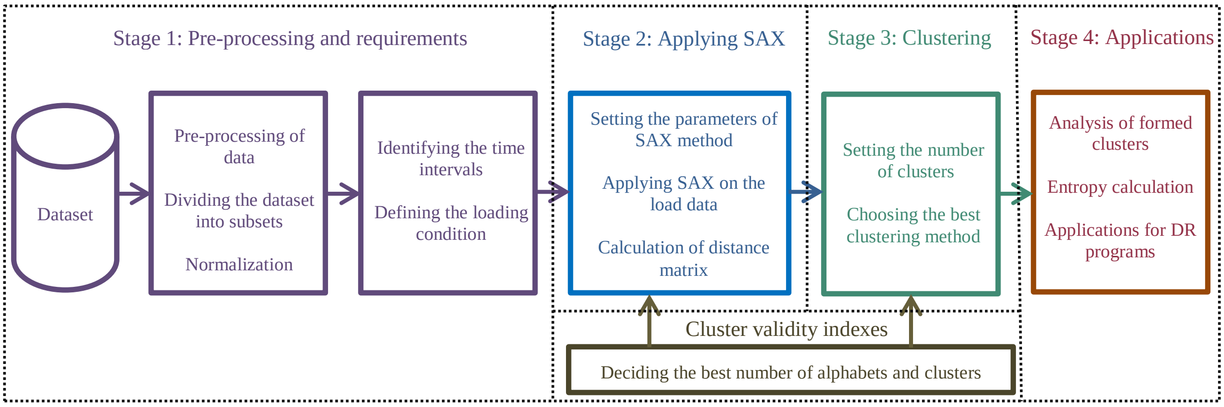
Stage 1: Pre-processing and requirements
Stage 2: Applying SAX
Stage 3: Clustering
Stage 4: Applications
Dataset
Pre-processing of data
Dividing the dataset into subsets
Normalization
Identifying the time intervals
Defining the loading condition
Setting the parameters of SAX method
Applying SAX on the load data
Calculation of distance matrix
Setting the number of clusters
Choosing the best clustering method
Analysis of formed clusters
Entropy calculation
Applications for DR programs
Cluster validity indexes
Deciding the best number of alphabets and clusters

Fig. 4. Stages of the methodology

- **Data preparation:** prior to data analysis, load data preparation is carried out in which the data set is divided into weekdays and weekend sets after the data cleaning process. Public holidays are also extracted from the weekdays and the customers' daily load profiles are normalized based on the daily maximum consumption. Normalization is needed since the direct use of raw data results in clusters that only reflect load magnitudes. On the other hand, normalized load patterns represent shape information and the variation of the load during the day [42]. Loading conditions are defined in this stage and the data are evaluated to determine the proper time intervals during the day, which are needed as the breakpoints on the time axis in the SAX method.
- **Implementing the SAX:** the second stage applies the SAX on the load patterns and calculates the distance matrix. Two modifications are applied in this stage to the SAX method to accommodate the underlying assumptions of residential consumption and to achieve better clustering results.
- **Clustering:** clustering using a hierarchical algorithm is conducted in Stage 3. Different hierarchical methods are evaluated and the best one is selected for the segmentation of load curves. Furthermore, the effects of SAX parameters and number of clusters are examined in Stages 2 and 3.
- **Applications and DR targeting:** The analysis of the results and further applications for DR programs are conducted in Stage 4. In this step, the entropy concept from the information theory is utilized to rank customers based on their stabilities over time which can be further used in applying customized DR programs for different customers. In other words, this measure ranks customers based on their variability in their DLCs over time and distinguishes between the customers with stable and regular behavior and customers with irregular and unstable behavior.

3. Methodology

In this section, the concepts and mathematical formulations of SAX technique, clustering algorithm, and entropy measure are illustrated.

3.1. SAX method

Various methods have been proposed in the data mining literature for the high level representation of time series data including the SAX technique. SAX is a symbolic representation for the time series data which discretizes the original data into symbolic strings. This technique, proposed by Lin et al. [26], uses an intermediate transformation of raw data called piecewise aggregate approximation (PAA) for the dimensionality reduction and then, symbolizes the PAA representation into a string composed of some alphabets. PAA partitions the time domain into a specified number of time frames and reduces the length of the original time series by replacing the data falling in the same time frame with their corresponding mean value. Usually, these time intervals have the same size. In our proposed method, we determine the time intervals based on the approach defined in Section 4.2. Mathematically expressing, PAA transforms the original time series $\mathrm{X} = \{x_1, x_2, \dots, x_H\}$ to a vector of $\mathrm{C} = \{c_1, c_2, \dots, c_K\}$ $(\mathrm{K} < \mathrm{H})$ in which c_i is defined as:

$$c_i = \frac{\sum_{x_j \in T_i} x_j}{k_i} \tag{1}$$

where T_i and k_i represent the ith time interval and the number of the data values falling in that time interval, respectively.

Once c_i is obtained, their SAX representation can be realized by defining a series of breakpoints $\{z_1, z_2, \dots, z_{Q-1}\}$ which partition the amplitude axis into Q intervals. These breakpoints can be determined based on the quantile of the statistical distribution that represents the probability density of the amplitudes in the whole data set [32]. A symbol is then assigned to each of these intervals and the PAA values are mapped to these symbols according to the interval they fall in. Therefore, if the set of symbols are defined as $\{\alpha_1, \alpha_2, \dots, \alpha_Q\}$ and $z_{j-1} \le c_i < z_j$, then c_i will be mapped to α_j. In this way, the original time series can be replaced by a SAX "word".

Fig. 5 demonstrates the process of generating the SAX word for a typical normalized load curve. Each day is divided into four intervals which have the same length of six and the amplitude breakpoints are selected as $\{0.25, 0.5, 0.75\}$. The data points falling in the same time frame are averaged and mapped to the letters $\{a, b, c, d\}$ hence, resulting in the SAX word *"abababacabbcaaacabab"*.

In the proposed method, every normalized daily load curve which consists of 48 data points will be replaced by its SAX representation of length five as a day is divided into five periods.

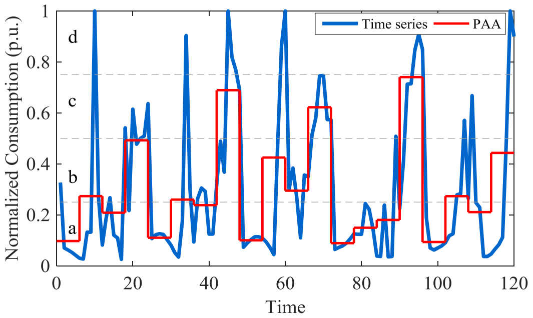

Fig. 5. PAA and SAX representation of a load curve for 5 days

3.2. Clustering stage

Considering that the SAX entries are categorical data, certain clustering algorithms such as K-means cannot be used for the clustering. Other algorithms including K-modes, hierarchical, and DBSCAN can be applied to partition the SAX representation of daily load curves into a number of clusters. Hierarchical agglomerative clustering is a good choice which is used in this paper.

The hierarchical algorithm produces a tree or dendrogram by either agglomerative (bottom-up) or divisive (top-down) methods. In almost all the studies in the power system domain, the agglomerative approach is used as the preferred hierarchical method. In the agglomerative method, initially each instance is classified as a cluster and then clusters are merged iteratively to build a bottom-up hierarchy of the clusters until just one cluster is obtained. This formed hierarchy can be cut at any given level to produce the final clusters.

To apply the method, firstly, the distance between the pairs of data points (load curves) is calculated and a similarity matrix is built. In order to apply the clustering algorithm on the SAX representations of the data, the distance between two SAX words should be defined. When dealing with the raw data, Euclidean distance is usually used to calculate the distance between two load curves or two RLPs. For the symbolic representation of the load curves a proper distance metric needs to be defined. When the time domain is partitioned into equal time intervals, the following distance definition known as MINIDIST is usually used for calculating the dissimilarity between two SAX words W_1 and W_2 [26]:

$$MINIDIST(W_1, W_2) = \sqrt{k.\sum_{i=1}^{K} (dist(\alpha_i, \beta_i))^2} \tag{2}$$

$$dist(\alpha, \beta) = \begin{cases} 0, & if\ |\alpha - \beta| \leq 1 \\ z_{\max(\alpha,\beta)-1} - z_{\min(\alpha,\beta)}, & otherwise \end{cases} \tag{3}$$

where α_i and β_i represent the ith symbols of W_1 and W_2, respectively and $k = H/K$ is the number of data points in each time interval.

Therefore, when calculating the distance between two SAX words, the distance between their individual symbols will be calculated first using (3). According to (3), if two symbols are adjacent, then the distance between them is zero, otherwise, the distance between them is calculated using the amplitude breakpoints. For instance, in the previous example, $dist(b, c) = 0$ and $dist(b, d) = 0.75 - 0.5 = 0.25$. It can be proved that this definition for the distance between SAX words lower-bounds the Euclidean distance between the two original time series.

We apply two modifications to the above distance calculation. Since the partitions of time axis have different sizes, the fixed k in (2) is replaced by k_i which is equal to the data points in each time interval. Accordingly, Eq. (2) should be re-written as:

$$MINIDIST(W_1, W_2) = \sqrt{\sum_{i=1}^{K} k_i.(dist(\alpha_i, \beta_i))^2} \tag{4}$$

The definition in (3) considers the distance of adjacent symbols as zero and ignores the minimum and maximum points of time series. Therefore, here, the second modification is proposed based on the work in [43] for calculation of the distance in (3), which shows better results for clustering purposes. Firstly, for each interval, an indicator can be defined as following:

$$Ind_i = \frac{z_{i-1} + z_i}{2} \tag{5}$$

For the lowest region, z_{i-1} is the global minimum and for the highest one z_i is the global maximum. Secondly, the distance between two intervals can be defined as the distance between their corresponding indicators:

$$dist(\alpha, \beta) = \left|Ind_\alpha - Ind_\beta\right| \tag{6}$$

Using the above definitions, the distance between the pairs of DLCs, which are represented by their SAX words, is calculated. This process results in a distance matrix which will be used as the input of the hierarchical clustering algorithm. Based on this matrix, clusters are merged using a linkage criterion. The linkage is an evaluation function which indicates the best candidates for merging. Likewise the dissimilarity measure, the choice of linkage can also have an impact on the final clustering outcomes.

The average linkage is selected for the hierarchical clustering since the obtained results confirmed its superiority to the other methods. In the average clustering method, the distance between two clusters is defined as the average dissimilarity between instances from two clusters. Therefore, in contrast with some other linkage criteria such as single and complete, it considers the similarity between all pairs of instances which belong to the clusters. Let C_p and C_q be two clusters with n_p and n_q members, respectively. Then, based on the average method the distance between these two clusters is:

$$D(C_p, C_q) = \frac{1}{n_p.n_q} \sum_{\substack{X_i \in C_p \\ Y_j \in C_q}} d(X_i, Y_j) = \frac{1}{n_p.n_q} \sum_{\substack{W_i \in C_p \\ W_j \in C_q}} MINIDIST(W_i, W_j) \tag{7}$$

3.3. Cluster validity indexes

The clustering results can be assessed using clustering validity indexes. Usually, several cluster validity indexes are used simultaneously and based on the results of all of them, the final decision is made. Davies-Bouldin Index (DBI) and Mean Index Adequacy (MIA) are among the most popular ones in the power system literature and therefore, they are used in this study too. The lower values of these indicators imply better clustering results. The definitions of these two indicators are given in the following:

$$DBI = \frac{1}{K} \sum_{i=1}^{K} max_{j \neq i} \left\{ \frac{\left[\frac{1}{n_i} \sum_{x \in \boldsymbol{C}_i} d(x, c_i) + \frac{1}{n_j} \sum_{x \in \boldsymbol{C}_j} d(x, c_j)\right]}{d(c_i, c_j)} \right\} \tag{8}$$

$$MIA = \sqrt{\frac{1}{K} \sum_{i=1}^{K} {d_{C_i}}^2} \tag{9}$$

where,

d_{C_i} = the distance between cluster center c_i and the members of the ith cluster $\boldsymbol{C}_i$= $\sqrt{\frac{1}{n_i}\sum_{x\in\boldsymbol{C}_i} d^2(x, c_i)}$

In these formulas, K, c_i, n_i are the number of clusters, center of $\boldsymbol{C}_i$, and the number of data points belonging to $\boldsymbol{C}_i$, respectively. $d(x, y)$ represents the distance between the observations x and y.

3.4. DR application

The consumption of a household can be characterized by several main features: total/average energy consumption [44] [45], load shape [3] [4], the amount and time of the peak [35], and the stability in usage pattern over time. The first three determinants are well studied in the literature. In spite of its importance, the stability of the customer's usage behavior over time has been addressed in very limited studies [37], [39]. Nonetheless, it remains as one of the main challenges of DR implementation. In the following, using the results of clustering in the last subsection, the customers will be ranked based on their stability over time.

By applying the clustering algorithm, the DLCs of each dwelling will be assigned to different clusters based on the customer's consumption habits. The higher variability of daily consumption patterns means that the DLCs will belong to the greater number of clusters. This gives an intuitive measure to compare households based on their stability over time. However, this method does not provide a fair basis for comparing customers. For example, the daily curves of two different customers can be assigned to the same number of clusters in spite of having completely different patterns as shown in Table 3. Here, there are 100 DLCs for each customer and the number of final clusters is set to 10. As can be observed, customer A has a very regular pattern since almost all of his/her daily curves are assigned to Cluster # 2, while customer B follows much diverse consumption habits as his/her daily curves are distributed among four different clusters.

Table 3 A sample assignment of daily curves of two customers to different clusters

	Cluster Number			
	#2	#5	#7	#10
Customer A	97	1	1	1
Customer B	23	27	19	31

Therefore, a more sophisticated metric is needed to quantify the variability of households and rank them. We borrow the notion of entropy form the information theory to classify customers based on their stability over time. Shannon entropy is a popular entropy measure in various fields and is used in this paper too. It can be expressed as [46]:

$$H = -\sum_{i=1}^{N} p_i \,.\, log_b(p_i) \tag{8}$$

where N is the number of classes, and p_i is the probability (relative frequency) of an object from the ith class appearing. b is the base of the logarithm and its common values are 2, e, and 10.

In our context, each class represents a cluster and the relative frequency, p_i, is the number of times the DLC of a household belongs to the cluster i. Therefore, the entropy metric not only considers the number of clusters but also the probability that they appear in the customers' DLCs. If a household has a completely stable pattern, which means that all the daily curves belong to just one cluster, the entropy will be 0 ($p_i = 1$). In other extreme case, if a customer follows a completely irregular pattern, which means that all the clusters are equally likely in his DLCs, the entropy will be the highest.

4. Preliminary analysis of the data set

4.1. Data set

The data used in this study are collected as a part of the smart metering trial project that was carried out by Commission for Energy Regulation (CER) in Ireland [47]. This data set contains the half-hourly consumption readings of more than 4000 residential customers for a period of one and half years. In this paper, the DLCs for a year of data starting from 1st of December 2009 are used for the analysis. An initial pre-processing of the data including the correction of missing values and the daylight saving time changes is carried out.

The analyses in this section and the next section are performed for weekday and weekend datasets.

R software and Matlab packages are used for the data preparation, application of the method, and visualizing the results.

4.2. Finding the time intervals

In this stage, the aim is to find time periods during a day which are distinguished based on the energy consumption levels and the local troughs and peaks. A detailed analysis is performed to distinguish the proper periods during the day. These time intervals are the characteristic time periods where certain consumption patterns can happen. These periods correspond with the typical periods of household activity and the changes in the energy usages. They will be used as the time breakpoints in the SAX method.

Fig. 6 and Fig. 7 show the box and whisker plots of the daily consumption data of all customers for the weekdays and weekends, respectively. For simplicity, outliers are not shown in these plots. The median value of the consumption in every hour is shown by the solid line in the middle of each box. The two ends of the box show the first quartile, q_1, and third quartile, q_3. The bottom and top whiskers are specified as $q_1 - 1.5 \times (q_3 - q_1)$ and $q_3 + 1.5 \times (q_3 - q_1)$ and limit the range outside the box which is shown by the dashed line.

Analysis of these figures reveals important information about the consumption magnitude and pattern during different seasons. In addition, a significant difference can be observed in the energy consumption patterns of weekdays and weekends.

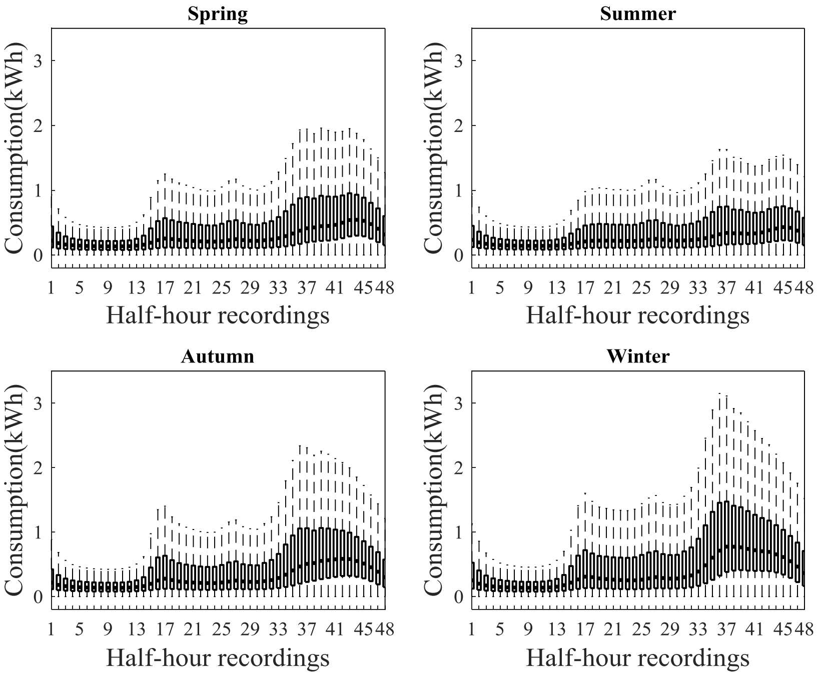

Fig. 6. Boxplots of weekday consumption in different seasons

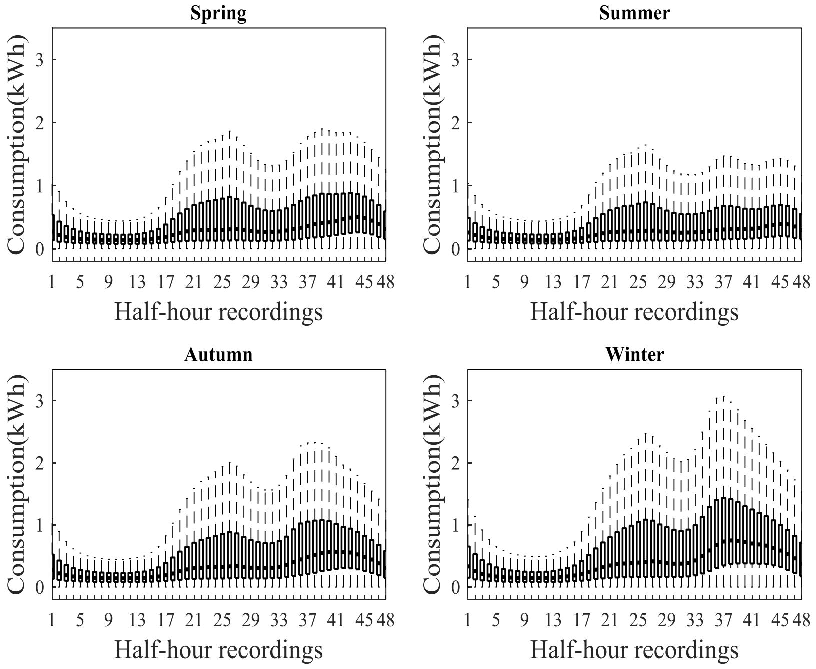

Fig. 7. Boxplots of weekend consumption in different seasons

For the weekdays, there are several local peaks which occur in the morning, mid-day, evening, and night periods. Although the occurring times of these peaks slightly differ among the seasons, the general trend is almost the same for all of them. As expected, the daytime peak during the weekends is much higher in comparison to weekdays. Specifically, for summer and spring, the daytime peak is the same or higher than the night peak. It is understandable since during the weekdays most of people are usually at workplaces, while they spend much more time at home during the weekends. There is no significant change in the night consumption. Another main difference for the weekends is the shifting of the local peak of the morning to mid-day. Such pattern is also expected since occupants usually wake up at later hours on Saturday and Sunday.

Table 4 reports the time intervals which are inferred from these plots which distinguish the periods of household activity and are used as the input of SAX method.

Table 4 Characteristic time intervals of the day (used for the SAX method)

Weekdays		
Recording number	**Hours**	**Period**
3-14	1 am- 7 am	Overnight 2
15-18	7 am- 9 am	Morning
19-33	9am- 4:30 pm	Daytime
34-45	4:30 pm- 10:30 pm	Evening
46-2	10:30 pm- 1 am	Overnight 1
Weekends		
Recording number	**Hours**	**Period**
3-16	1 am- 8 am	Overnight 2
17-20	8am- 10 am	Morning
21-34	10 am- 5 pm	Daytime
35-46	5 pm- 11 pm	Evening
47-2	11 pm- 1 am	Overnight 1

5. Case study

5.1. Application of SAX and clustering algorithms

The time periods that were defined for the SAX method cover the main time intervals during the whole year. However, the usage patterns of customers might slightly vary based on the temperature and seasonal changes. Usually, more similar consumption habits happen during a season. The current practice in the literature is to divide a year into 4 (or sometimes 5) seasons and study the consumption behavior based on the seasons[12] [28] [42]. In our case, since we were studying the customer stability over time, we tried to define the season based on the temperature variations to expand the time period of study compared to the usual practice which considers only one pre-defined season of data. The analysis indicates that there is a good correlation between the temperature values (from Irish Meteorological Service [48]) and the overall consumption pattern as shown in Fig. 8. This approach represents a more realistic scenario as well as more number of days, resulting in 93 working days including 11 days from the end of spring, 18 days from the beginning of autumn, and all the working days of summer (64 days).

The presented concepts are firstly demonstrated using a dataset comprising of 300 customers (Sections 5.1 to 5.4), while a much larger dataset consisting of 4141 customers is used for further DR analysis (section 5.5). Therefore, 27,900 DLCs (300 customers × 93 days) are available for this loading condition. For the weekends, a year of data is considered which consists of 104 days. Consequently, for the time period under study, the total number of 31200 DLCs will be available for the weekends.

If the occupants are not at home for some days, the household’s consumption will be very low and almost the same for the whole day. So, after normalization, all the values will be around 1 and this will result in a consumption pattern that seems stable over days. Therefore, as an additional condition, only those customers whose usage in each day of the period under study is more than 3 kWh are selected.

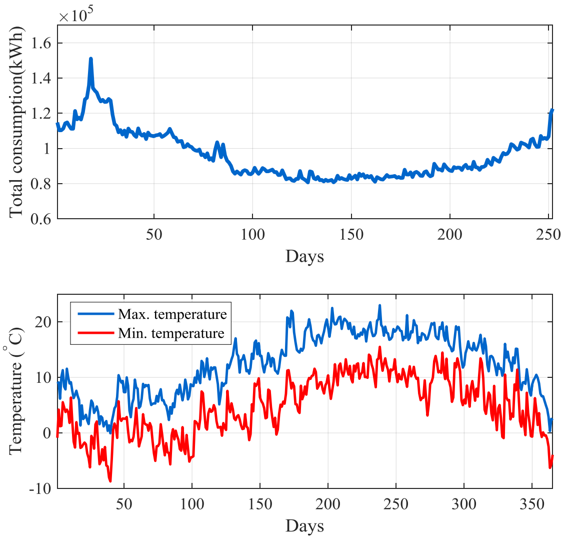

Fig. 8. Daily total consumption and temperature variation

The analysis is performed for various conditions mainly by selecting different number of symbols (amplitude partitions) and different number of clusters. As will be explained in the next subsections, using a seven-alphabet shows superior clustering performance and hence, it is used for explaining the results.

The amplitude axis is partitioned into seven regions which are defined using the quantile of the PAA values. Accordingly, there are six breakpoints which are determined based on the cumulative density function (CDF) of the whole data. It means that these breakpoints divide the CDF curve into seven equiprobable regions as shown in Fig. 9.

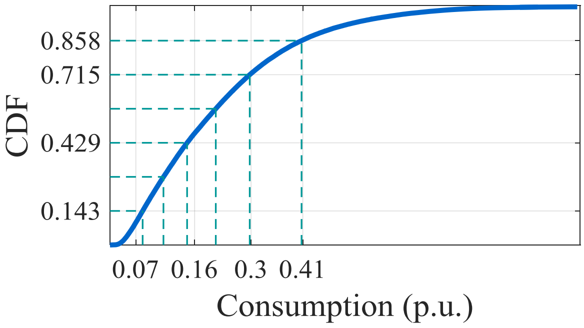

Fig. 9. CDF of the whole data

These regions correspond to the letters ‘a’ to ‘g’. The normalized daily load shapes are transformed into their corresponding SAX words. The distance matrix is constructed using the modified MINIDIST function defined in (4)-(6). Finally, the SAX words are clustered using the hierarchical clustering algorithms into clusters. *Cophenetic* correlation coefficients are calculated for average, single, complete, and Ward clustering methods in which the results demonstrate the better performance for average clustering. Finally, the entropy of each customer is calculated based on the frequency of appearance of clusters in his DLCs.

It should be noted that the maximum number of different SAX words cannot be more than $7^5 = 16708$ as seven possible symbols exist for each period. Consequently, even if the number of daily load shapes increases considerably, the possible combinations of the symbols “a” to “g” are limited to this number.

5.2. Analysis of weekday and weekend clusters

The number of clusters is determined based on the values of DBI and MIA indexes. Following the analysis presented in Section 5.4, the number of clusters is set to 120. Analysis of the formed clusters and their shapes can give valuable information about the consumption patterns. Fig. 10 depicts the clustering results for the weekday dataset in which the centers for 30 clusters with the largest number of members are shown. The center is calculated by averaging all the curves that belong to the cluster.

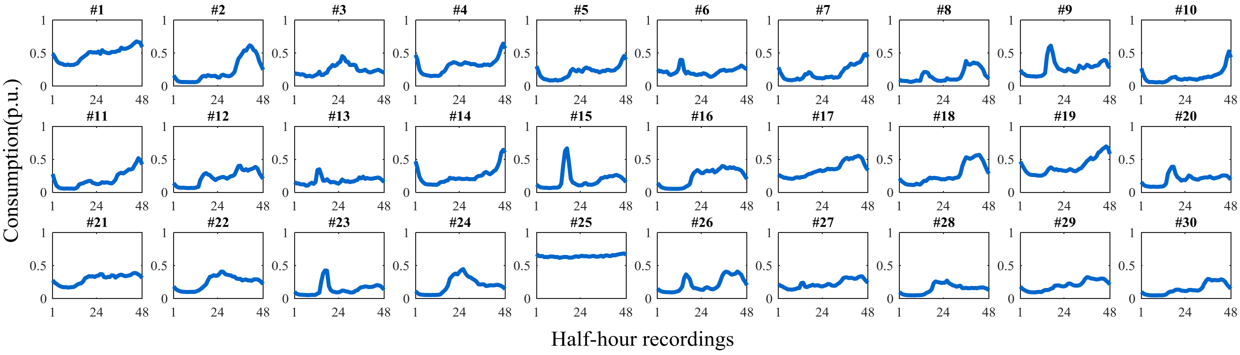

Fig. 10. Centers of the clusters with the highest number of DLCs for the weekday data set

Each cluster characterizes a different consumption pattern. Especially, these patterns are visible:

- Morning peak, clusters #6, #9, #13, and #15, #20, and #23: except for cluster # 6 whose peak happens at early morning (around 6), other clusters show a late peak during the morning. Clusters # 9 and # 15 represent a major peak which shows the energy consumption by various appliances in this period.
- Mid-day peak; clusters #3, #22, #24, and #28: for this category, the peak happens at early afternoon and, except for cluster # 3, they generally follow the same pattern of an extended peak. It means that the consumption slowly increases until the peak point and then again follows a slow decrease pattern.
- Late night peak; clusters #1, #4, #5, #7, #10, #11, #14, and #19: these clusters show a peak near to the midnight. Clusters #1, #7, #11, and #19 show a smooth steady increase in the usage from early afternoon till midnight compared with clusters #10 and #14 which have a sharper rise of consumption.
- Night peak; clusters #2, #17, #18, #27, #29, and #30: compared with other clusters in this category, clusters #2 and #18 display a significant increase in consumption during night time.
- Dual peaks; clusters #8, #12, #26: these clusters show dual peaks happening in the morning and in the evening.

Clusters #16, #21, and #25 classify those days when customer had an almost stable consumption during the day starting from the morning and lasting until midnight.

Clusters #29 and #30 have the largest number of members (4187 and 2406 DLCs, respectively) which account for around 24 per cent of all daily curves. Not surprisingly, they represent a consumption pattern similar to the general trend of energy use in summer which was shown in Fig. 6.

The same analysis that is performed for the week days can also be applied on weekend dataset. The centers of main clusters, that contain the largest number of members, are depicted in Fig. 11. It can be seen that weekend clusters generally have a peak in the afternoon and nights rather than in the morning. Especially, clusters #25 to #30 which have the largest number of members follow this pattern. Cluster #30 has 5541 members and its shape resembles the consumption patterns shown in Fig. 7.

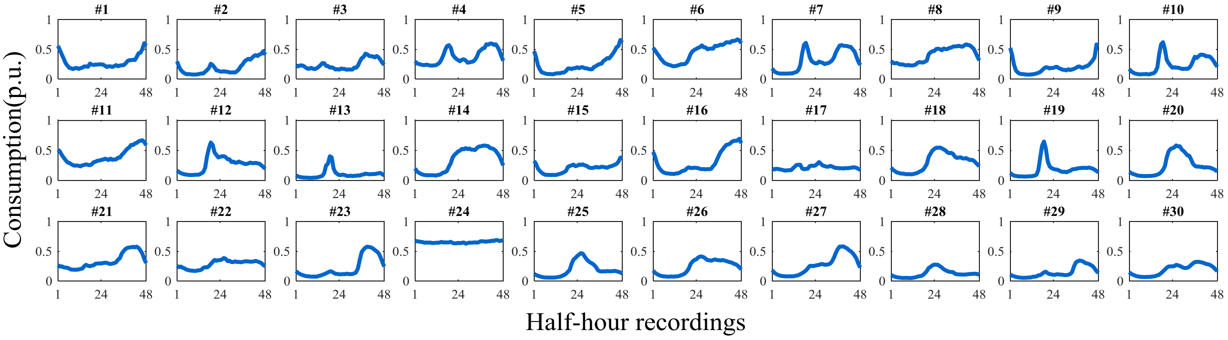

Fig. 11. Centers of the clusters with the highest number of DLCs for the weekend data set

5.3. Entropy analysis

The results of the method distinguish the customers based on their variability in the defined periods. The highest entropy that a customer can have is $-\log\left(\frac{1}{93}\right) = 1.9685$ which happens when each DLC belongs to a different cluster. Here, the DLCs of the dwelling with the lowest entropy are assigned to just 4 clusters, while for the customer with the highest entropy they belong to more than 50 clusters. The entropy values for the former and latter cases are 0.2444 and 1.5991, respectively. Fig. 12 shows the daily curves of two customers with stable consumption behavior and two customers with variable consumption behavior that are decided based on the entropy analysis. As can be seen, even the stable customers show slightly different patterns from one day to another day which is an inherent feature of the residential energy usage.

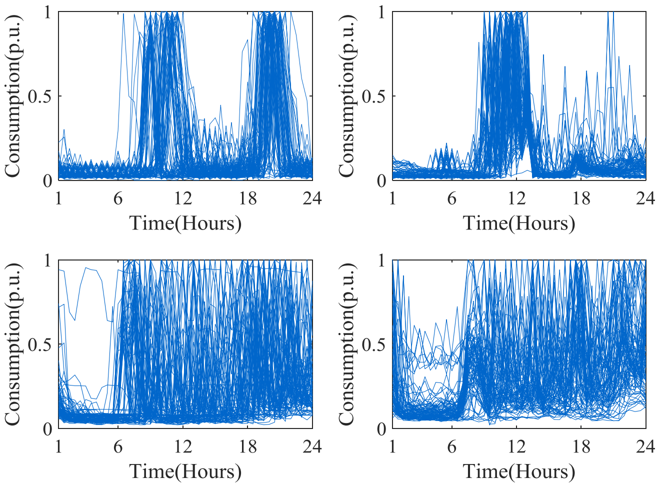

Fig. 12. Load curves of two customers with stable consumption behavior (top) and two customers with variable consumption behavior (bottom)

Entropy analysis reveals that the weekend consumption is more variable than the weekdays, which can be associated with the more stable schedule of households during the weekdays. Furthermore, the study shows that those customers who show a stable behavior on the weekdays do not necessarily follow a regular pattern on the weekends.

5.4. Effect of amplitude partitioning and number of clusters

The final clustering results are affected by the number of regions of amplitude axis i.e. the number of alphabets. To investigate this effect on clustering results, the number of amplitude breakpoints has been changed from 4 (5 regions) to 7 (8 regions). On the other hand, the effect of the number of clusters is also studied by changing it from 50 to 300.

Fig. 13 and Fig. 14 display the results for DBI and MIA, respectively. It can be seen that, both DBI and MIA values decrease with the increase of the number of clusters. However, they do not change significantly after around 120 and for this reason, the number of clusters is set to 120. In addition, it can be observed that the best clustering has been achieved when the alphabet of size 7 is used.

It is also interesting to study the effect of these variables on entropy calculations and their ability to distinguish stable customers. To this end, for each number of clusters, the entropies of customers are calculated for a different number of alphabets. Then, 100 customers with the lowest entropies for each of these four scenarios and for the different number of clusters were obtained. It is observed that there are at least 66 common users among the 100 users for different number of alphabets.

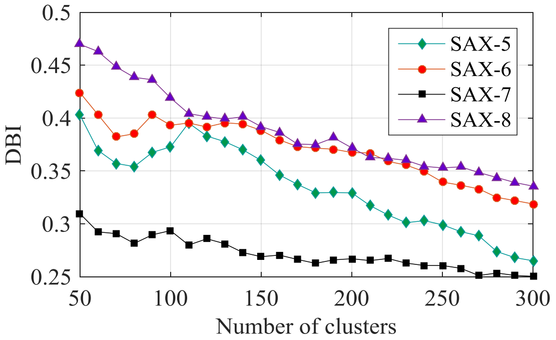

Fig. 13. DBI values for different size of alphabets and different number of clusters

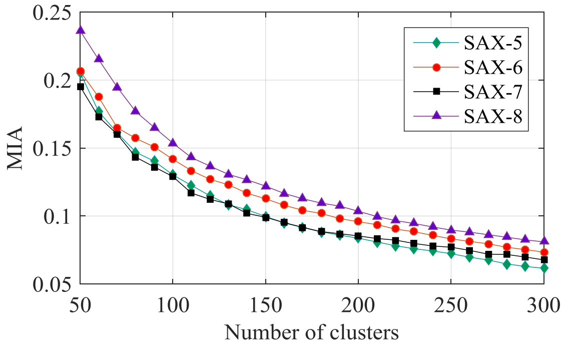

Fig. 14. MIA values for different size of alphabets and different number of clusters

5.5. Entropy analysis for a large number of customers

The presented method in the last section is precise, however, with the increase of the number of DLCs the computational time increases. The main reason is that SAX words are categorical variables and the distance matrix for the clustering algorithm needs to be constructed. Consequently, if there are n SAX words, the order of distance matrix is $n \times n$. More sophisticated approaches for clustering are needed that is beyond the scope of this study. Therefore, here, we adopt a more straightforward approach in order to be able to compare a large number of customers. The steps of the proposed method are illustrated in the following:

- Step 1: Define the *K* characteristic load shapes based on which all the load curves can be compared. These characteristic load curves can be defined by the user or for example, as a preliminary stage can be found by a simple clustering algorithm like K-means.
- Step 2: Convert these *K* characteristic load shapes to SAX words (representative SAX words).
- Step 3: Convert all the DLCs of each customer to SAX words.
- Step 4: Compute the distance of each DLC of the customer with the representative SAX words and assign it to the most similar one.
- Step 5: Calculate entropy.

The case studies for the working days and weekend days are illustrated in the following.

The simulations are carried out for the working days of one year which comprises 252 days. Again, only those customers with the daily usages above 3 kWh are selected for the analysis. The partitioning of time and amplitude axis is the same as the previous section. The number of representative load shapes is decided large enough to represent various load patterns. It should be noted that some of the representative SAX words of these K centers might be identical and hence, the replicated ones need to be removed. In our analysis, the final number of representative SAX words is decreased to 39. For each DLC, the distance of its SAX word with these 39 load shapes is computed and it is assigned to the most similar load shape. Finally, the entropy of each customer is calculated.

The histogram in Fig. 15 shows how the DLCs of customers are assigned to different clusters. It can be observed that the DLCs of most of the customers are assigned to around 20 clusters. The DLCs of the most stable customer belong to only 2 clusters, while the ones of the most variable customer are assigned to 32 clusters. However, as emphasized before, this measure cannot independently reveal the customer's variability. Fig. 16 shows the load shapes of some of the stable customers.

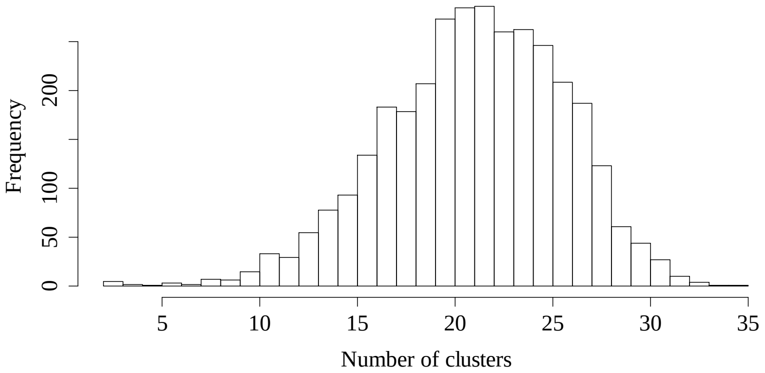

Fig. 15. Cluster assignment distribution

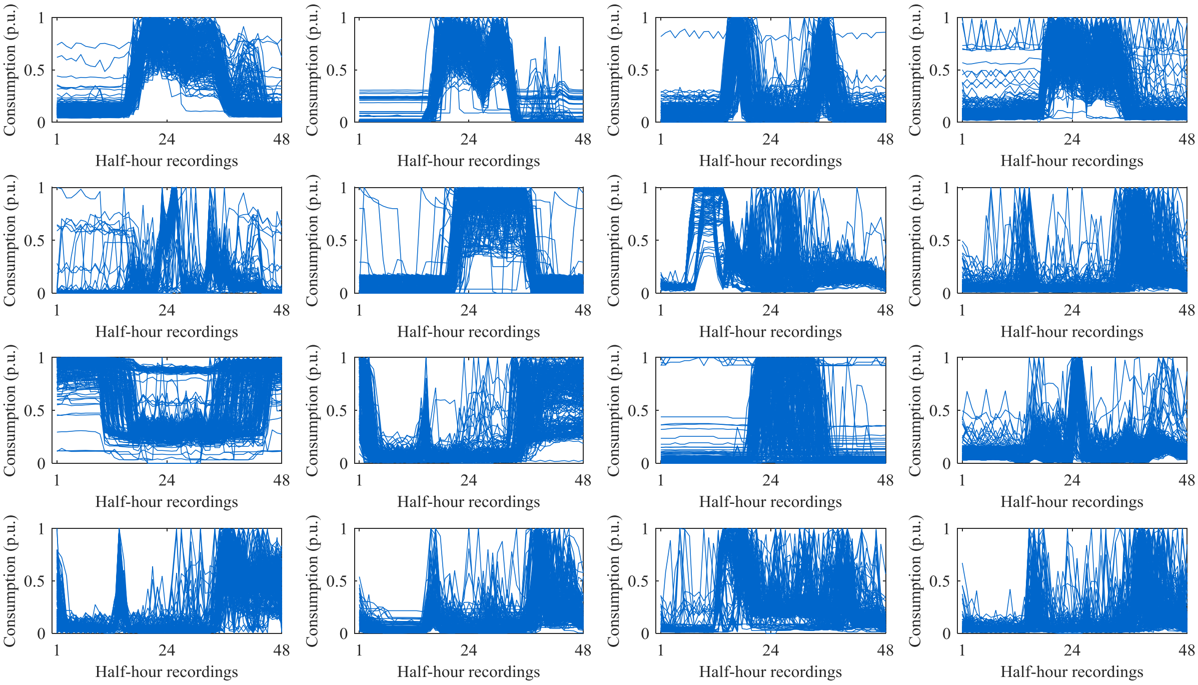

Fig. 16. Load curves of sample customers with stable consumption behavior for weekday dataset

There are 104 days for the weekend dataset and the representative load patterns are different for this dataset. The entropy analysis for weekend days reveals that customers show more variable consumption patterns during the weekends. Fig. 17 shows the sample stable customers for the weekend dataset.

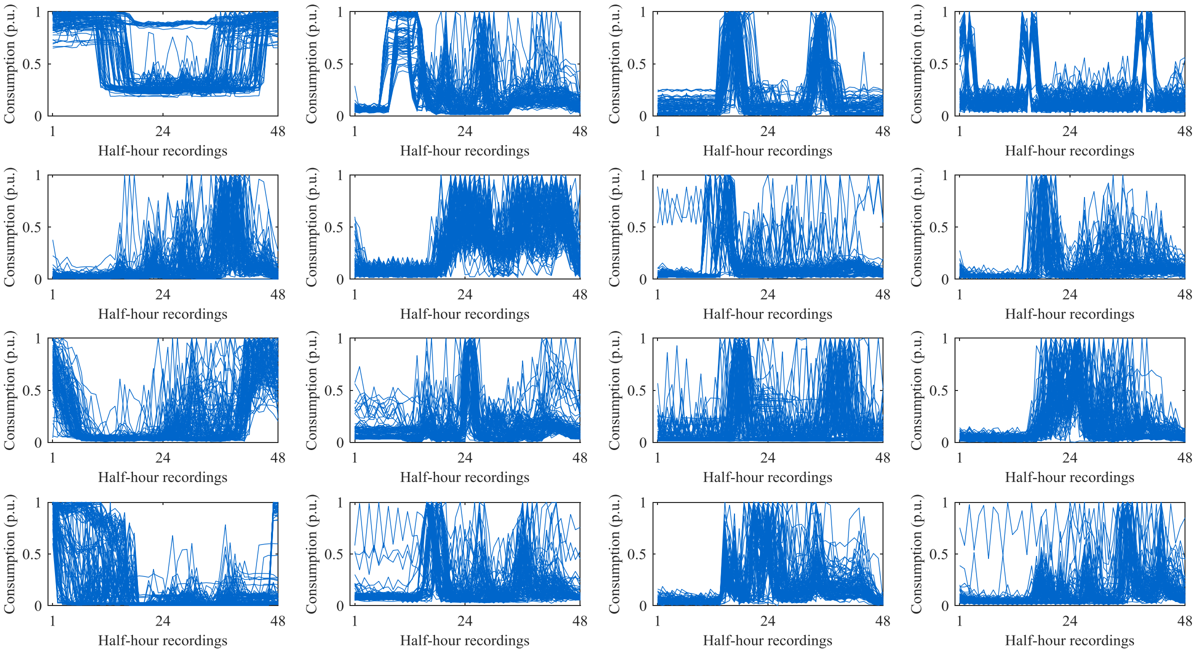

Fig. 17. Load curves of sample customers with stable consumption behavior for weekend dataset

In the next step, we investigate the stable customers who are common between the two datasets. To this end, the customers are ranked based on their entropies and the top common stable customers in the two datasets are found. The analysis shows that there is not any direct relationship between the stable customers of the weekday dataset and the weekend dataset. For example, among the top 500 stable customers for each of weekday and weekend dataset, only 74 customers are common.

6. Conclusions

Future smart grids will be equipped with monitoring devices such as smart meters which collect the massive volume of consumption data of electricity users at frequent intervals. Data mining of these energy data can benefit stakeholders in power systems in various ways, for example, for implementing proper DR settlements. In this paper, using a combination of SAX technique as a data size reduction method and hierarchical clustering algorithm, DLCs of customers were segregated into certain clusters. Also, the stability of consumption pattern was investigated using the entropy concept.

The results of clustering and entropy calculation provide insights for offering DR programs to the electricity customers. First of all, the major DLCs which represent the main consumption habits can be determined as shown in figures 10 and 11. This, in turn, will help the utilities to build DR programs or customized time of use tariff structures based on the load patterns. Furthermore, the entropy analysis enables them to divide their customers to various groups which can be targeted differently. For example, low entropy customers whose load peaks happen at the same period as system peak are good candidates for price-based DR. On the other hand, variable users can be targeted by suitable incentive-based DR.

The case studies using a large number of daily load patterns demonstrate the applicability of the proposed method for DR management. Especially, the method is able to capture the major consumption patterns of the households and overcome the problems of the previous studies [37], which mainly dealt with the original DLCs of residential customers that varies a lot on the daily and seasonal bases. Use of SAX method can produce superior results, since the residential consumption patterns, which inherently display a lot of variability, are transformed to a limited number of SAX words.

Acknowledgments

This research is supported by an Australian Government Research Training Program Scholarship and is supported in part by the ARCLab facility at University of Technology Sydney.

References

[1] Leiva J, Palacios A, Aguado JA. Smart metering trends, implications and necessities: A policy review. Renewable and Sustainable Energy Reviews. 2016;55:227-33.
[2] Fu X, Zeng X-J, Feng P, Cai X. Clustering-based short-term load forecasting for residential electricity under the increasing-block pricing tariffs in China. Energy. 2018;165:76-89.
[3] Chicco G. Overview and performance assessment of the clustering methods for electrical load pattern grouping. Energy. 2012;42:68-80.
[4] Tsekouras G, Kotoulas P, Tsirekis C, Dialynas E, Hatziargyriou N. A pattern recognition methodology for evaluation of load profiles and typical days of large electricity customers. Electric Power Systems Research. 2008;78:1494-510.
[5] Ozawa A, Furusato R, Yoshida Y. Determining the relationship between a household's lifestyle and its electricity consumption in Japan by analyzing measured electric load profiles. Energy and Buildings. 2016;119:200-10.

[6] Pereira R, Fagundes A, Melício R, Mendes VMF, Figueiredo J, Martins J, et al. A fuzzy clustering approach to a demand response model. International Journal of Electrical Power & Energy Systems. 2016;81:184-92.

[7] Biscarri F, Monedero I, García A, Guerrero JI, León C. Electricity clustering framework for automatic classification of customer loads. Expert Systems with Applications. 2017;86:54-63.

[8] Zhang T, Zhang G, Lu J, Feng X, Yang W. A new index and classification approach for load pattern analysis of large electricity customers. IEEE Transactions on Power Systems. 2012;27:153-60.

[9] Räsänen T, Ruuskanen J, Kolehmainen M. Reducing energy consumption by using self-organizing maps to create more personalized electricity use information. Applied Energy. 2008;85:830-40.

[10] Crow M. Clustering-based methodology for optimal residential time of use design structure. North American Power Symposium (NAPS), 2014: IEEE; 2014. p. 1-6.

[11] Haben S, Singleton C, Grindrod P. Analysis and clustering of residential customers energy behavioral demand using smart meter data. IEEE Transactions on Smart Grid. 2016;7:136-44.

[12] Wang F, Li K, Duić N, Mi Z, Hodge B-M, Shafie-khah M, et al. Association rule mining based quantitative analysis approach of household characteristics impacts on residential electricity consumption patterns. Energy conversion and management. 2018;171:839-54.

[13] Yang J, Zhao J, Wen F, Dong Z. A Model of Customizing Electricity Retail Prices Based on Load Profile Clustering Analysis. IEEE Transactions on Smart Grid. 2018.

[14] Fahiman F, Erfani SM, Rajasegarar S, Palaniswami M, Leckie C. Improving load forecasting based on deep learning and K-shape clustering. Neural Networks (IJCNN), 2017 International Joint Conference on: IEEE; 2017. p. 4134-41.

[15] Benítez I, Díez J-L, Quijano A, Delgado I. Dynamic clustering of residential electricity consumption time series data based on Hausdorff distance. Electric Power Systems Research. 2016;140:517-26.

[16] Benítez I, Quijano A, Díez J-L, Delgado I. Dynamic clustering segmentation applied to load profiles of energy consumption from Spanish customers. International Journal of Electrical Power & Energy Systems. 2014;55:437-48.

[17] Pal R, Chelmis C, Frincu M, Prasanna V. Time Series Clustering for Demand Response, An Online Algorithmic Approach. Online: wwwcsuscedu/assets/007/93954pdf.

[18] Varga ED, Beretka SF, Noce C, Sapienza G. Robust real-time load profile encoding and classification framework for efficient power systems operation. IEEE Transactions on Power Systems. 2015;30:1897-904.

[19] Carpaneto E, Chicco G, Napoli R, Scutariu M. Electricity customer classification using frequency–domain load pattern data. International Journal of Electrical Power & Energy Systems. 2006;28:13-20.

[20] Xiao Y, Yang J, Que H, Li MJ, Gao Q. Application of wavelet-based clustering approach to load profiling on AMI measurements. Electricity Distribution (CICED), 2014 China International Conference on: IEEE; 2014. p. 1537-40.

[21] Zhong S, Tam K-S. Hierarchical classification of load profiles based on their characteristic attributes in frequency domain. IEEE Transactions on Power Systems. 2015;30:2434-41.

[22] Räsänen T, Kolehmainen M. Feature-based clustering for electricity use time series data. International Conference on Adaptive and Natural Computing Algorithms: Springer; 2009. p. 401-12.

[23] Motlagh O, Foliente G, Grozev G. Knowledge-mining the Australian smart grid smart city data: A statistical-neural approach to demand-response analysis. Planning Support Systems and Smart Cities: Springer; 2015. p. 189-207.

[24] Abreu JM, Câmara Pereira F, Ferrão P. Using pattern recognition to identify habitual behavior in residential electricity consumption. Energy and Buildings. 2012;49:479-87.

[25] Lam JC, Wan KKW, Cheung KL, Yang L. Principal component analysis of electricity use in office buildings. Energy and Buildings. 2008;40:828-36.

[26] Lin J, Keogh E, Lonardi S, Chiu B. A symbolic representation of time series, with implications for streaming algorithms. Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery: ACM; 2003. p. 2-11.

[27] McLoughlin F, Duffy A, Conlon M. A clustering approach to domestic electricity load profile characterisation using smart metering data. Applied energy. 2015;141:190-9.

[28] Rhodes JD, Cole WJ, Upshaw CR, Edgar TF, Webber ME. Clustering analysis of residential electricity demand profiles. Applied Energy. 2014;135:461-71.

[29] Viegas JL, Vieira SM, Melício R, Mendes V, Sousa JM. Classification of new electricity customers based on surveys and smart metering data. Energy. 2016;107:804-17.

[30] Alzate C, Sinn M. Improved electricity load forecasting via kernel spectral clustering of smart meters. Data Mining (ICDM), 2013 IEEE 13th International Conference on: IEEE; 2013. p. 943-8.

[31] Koivisto M, Heine P, Mellin I, Lehtonen M. Clustering of connection points and load modeling in distribution systems. IEEE Transactions on Power Systems. 2013;28:1255-65.

[32] Notaristefano A, Chicco G, Piglione F. Data size reduction with symbolic aggregate approximation for electrical load pattern grouping. IET Generation, Transmission & Distribution. 2013;7:108-17.

[33] Assessment of demand-response and advanced metering. Federal Energy Regulatory Commission (FERC); Feb 2011.

[34] Rajabi A, Li L, Zhang J, Zhu J. Aggregation of small loads for demand response programs—Implementation and challenges: A review. Environment and Electrical Engineering and 2017 IEEE Industrial and Commercial Power Systems Europe (EEEIC/I&CPS Europe), 2017 IEEE International Conference on: IEEE; 2017. p. 1-6.

[35] Smith BA, Wong J, Rajagopal R. A simple way to use interval data to segment residential customers for energy efficiency and demand response program targeting. ACEEE Summer Study on Energy Efficiency in Buildings2012. p. 374-86.

[36] Labeeuw W, Stragier J, Deconinck G. Potential of active demand reduction with residential wet appliances: A case study for Belgium. IEEE Transactions on Smart Grid. 2015;6:315-23.

[37] Kwac J, Flora J, Rajagopal R. Household energy consumption segmentation using hourly data. IEEE Transactions on Smart Grid. 2014;5:420-30.

[38] Ibrahim C, Mougharbel I, Daher NA, Kanaan HY, Saad M, Georges S. An optimal approach for offering multiple demand response programs over a power distribution network. IECON 2018-44th Annual Conference of the IEEE Industrial Electronics Society: IEEE; 2018. p. 95-102.

[39] Albert A, Rajagopal R. Smart meter driven segmentation: What your consumption says about you. IEEE Transactions on power systems. 2013;28:4019-30.

[40] Samad T, Koch E, Stluka P. Automated demand response for smart buildings and microgrids: The state of the practice and research challenges. Proceedings of the IEEE. 2016;104:726-44.

[41] Rahiman FA, Zeineldin HH, Khadkikar V, Kennedy SW, Pandi VR. Demand Response Mismatch (DRM): Concept, Impact Analysis, and Solution. IEEE Transactions on Smart Grid. 2014;5:1734-43.

[42] Li R, Gu C, Li F, Shaddick G, Dale M. Development of low voltage network templates—Part I: Substation clustering and classification. IEEE Transactions on Power Systems. 2015;30:3036-44.

[43] Aghabozorgi S, Wah TY. Clustering of large time series datasets. Intelligent Data Analysis. 2014;18:793-817.

[44] Huebner G, Shipworth D, Hamilton I, Chalabi Z, Oreszczyn T. Understanding electricity consumption: A comparative contribution of building factors, socio-demographics, appliances, behaviours and attitudes. Applied energy. 2016;177:692-702.

[45] Kavousian A, Rajagopal R, Fischer M. Determinants of residential electricity consumption: Using smart meter data to examine the effect of climate, building characteristics, appliance stock, and occupants' behavior. Energy. 2013;55:184-94.

[46] Shannon CE. A mathematical theory of communication. Bell Syst Tech J. 1948;27:623-56.

[47] Commission for Energy Regulation (CER)- Electricity Smart Metering Customer Behaviour Trials (CBT)-Findings Report [Online] http://www.cer.ie/docs/000340/cer11080(a)(i).pdf.

[48] Met Éireann Data from Irish Meteorological Service, online: https://www.met.ie/climate/available-data/historical-data.