Load Profile Segmentation for Electricity Market Settlement

Murat Gunsay
Business Administration
Istanbul Sabahattin Zaim University
Istanbul, Turkey
mehmet.gunsay@std.izu.edu.tr

Canser Bilir
Industrial Engineering
Istanbul Sabahattin Zaim University
Istanbul, Turkey
canser.bilir@izu.edu.tr

Gokturk Poyrazoglu
Electrical & Electronics Engineering
Ozyegin University
Istanbul, Turkey
gokturk.poyrazoglu@ozyegin.edu.tr

***Abstract*— An unsupervised learning method is used to create clusters for electricity load profiles within a group of real customers. A time-series analysis method (hierarchical clustering) is adopted. A case study is conducted with real consumption data from residential, commercial, and industrial consumers to show the effectiveness of the proposed clustering method for load profiling. After the data cleansing, filtering, and normalization processes, the input dataset is divided into several clusters based on their profile differences. Later, various results are obtained to reflect different consumption patterns within a profile group by the selected distance measurement methods such as Euclidean and Dynamic Time Warping. The results obtained in the case study show that the proposed mathematical algorithm can be used to create realistic and scalable profiling subgroups (with percentages of similar consumptions in each cluster) instead of the traditional methods which cluster all profiles in a single big cluster. The proposed algorithm is used for a case study of Turkey; however, this study is adaptable to other European markets.**

Keywords— load profiling, clustering, consumption, market settlement

I. INTRODUCTION

The electricity load profile shows how a customer consumes electricity over time. Electricity consumption can be measured and the data can be stored and transferred by smart meters. The hourly load can represent the customer's consumption pattern. There are five main load profiling groups in Turkey including residential, industrial, commercial, agricultural irrigation, and street lighting. Although the consumption pattern of each group member is similar, there might be distinct differences within the groups themselves. Furthermore, even the subgroupings can have distinct differences. Most of the distribution system operators (DSO) use a single consumption profile for each group. However, a single profile can only cover a limited percentage of real individual consumptions. Hence, this study aims to identify different profiles (cluster) within each group that may cover a higher percentage of consumers.

The electricity consumption data from a smart meter is a set of numeric values following each other in a one-hour interval, and as such, can be considered a time-series dataset. Analysis of time series data especially in statistics and econometrics has been around for a long time. In parallel with the increase of computing abilities of computers and the development of data mining techniques, many studies have been conducted using time series analysis to provide the above-mentioned benefits in electrical systems [1–3]. In literature, many studies can be found where different countries' electric consumption data has been analyzed using clustering or grouping of load profiles [4, 5]. Although there are many known clustering methods, the main methods mentioned in the literature are hierarchical clustering, k-means, fuzzy k-means, follow-the-leader, and self-organizing maps (SOM). In this study, mainly, data collected from 3 different electric distribution companies were investigated. The real consumption data from around a thousand smart meters are collected. Daily and hourly electricity consumption data are analyzed using hierarchical clustering techniques with EEuclidian distance and Dynamic Time Warping (DTW) distance methods. Different linkage methods such as average linkage and Ward's linkage are used.

The rest of the paper is structured accordingly. Section II provides the details of data preparation including data collection, data cleansing, data filtering, and data normalization. Section III explains the clustering methods in time series analysis and distance measure methods such as Euclidean and dynamic time warping. Section IV gives the results of case studies conducted for a customer group in Turkey.

II. CLUSTERING

Hierarchical clustering (HC) is one of the most used classical methods for clustering load profiles [2, 4]. This method is based on a foundation called a dendrogram that has a (reversed) tree-like structure. The HC method has two approaches: a top-down approach and a bottom-up approach. Since the top-down approach (which is also called the divisive approach) requires a lot of computing power [8], the bottom-up approach is the preferred approach. In the bottom-up approach, each 24-hour load profile table is initially considered a cluster on its own. Therefore, a cluster number of "1" and a total load profile number of N are formed. The number of clusters to be used in the HC operation can be determined by looking at the dendrogram and the rate represented in each cluster [9]. As you go up, each cluster is combined with another cluster closest to it, and this process continues until a single cluster remains. From this aspect, the HC method can be called an agglomerative method, since all clusters become part of a larger stack until one large stack remains [9]. A sample dendrogram is given in Fig. 1 for a data set that occurs after a 24-hour electricity consumption data is normalized and transformed into a load profile for a year.

The dendrogram visually reveals the clustering relationship of similar clusters as well as outliers. In Fig. 1, the y-coordinate is the similarity coefficient and the x-coordinate is the total number of load profiles in the data set. The similarity of the clusters is negatively proportional to the y-coordinate value. Therefore, the similarity of the clusters C and D showed in Fig. 1, is more than the similarity of clusters A and B. Since the horizontal red line in the figure cut 6 vertical lines in total, it is possible to show the data set in 6 separate sets for a similarity coefficient of about 1. The population of load profile data under both clusters C and D

978-1-7281-6919-4/20/$31.00 ©2020 IEEE

appears to be more than the population of load profile data in 3 sub-clusters under cluster B. Dendrograms does work to visually show the distribution of clusters, but they are not used to determine the optimal number of clusters.

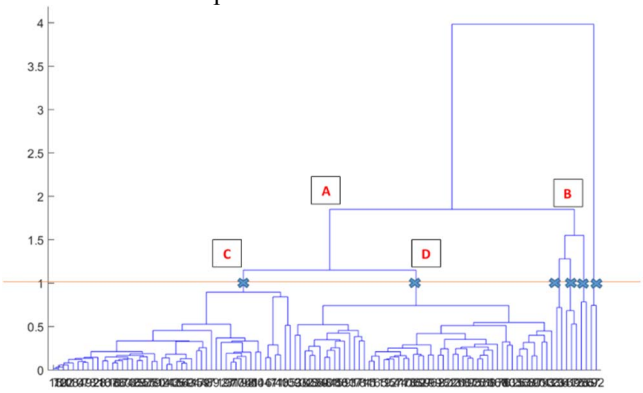

Fig. 1 A resultant dendrogram of hierarchical clustering algorithm

1) Distance Criteria

Many "Distance Criteria" are available to calculate the distance of the vectors (daily load profiles) in the data set. The most basic and best known of these criteria is "Euclidian Distance". There are many distance criteria other than Euclidean distance. The most well-known of these are Minkowski, Manhattan, and Mahalanobis distance. The Euclidean distance is the linear distance between points *P* and *Q* and it is represented by the Pythagorean formula given in (1), where *n* is the dimension of the vector *P* and *Q*.

$$P = (p_1, p_2, \dots, p_n) \quad Q = (q_1, q_2, \dots, q_n)$$
$$\sqrt{(p_1 - q_1)^2 + \cdots + (p_n - q_n)^2} = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2} \qquad (1)$$

2) Linkage Criteria

Linkage is an evaluation function [4] used in the hierarchical clustering method, after determining the distance criteria and forming the similarity matrix. The distance information combines the two closest distances to form a larger set. This binary operation continues until a single large cluster remains.

Some of the popular linkage methods are single-link, complete-link, average-link, and Ward's link. The single-link criterion is based on determining the similarity of two clusters by finding the distance between the closest members/elements of the two cluster sets [4, 10]. On the other hand, the complete-link criterion is based on determining the distance between two distant members/elements of the two cluster sets to find the similarity of the two clusters [4, 10]. The average-link criterion is based on determining the average distance between each member/element of one cluster with all the other members/elements of another cluster [4]. The common denominator of each of the three linkage criteria is that they all use the proximity matrix as input [10].

Ward introduced another linkage criterion to the literature named after his name [10]. According to this method, the similarity between the clusters is inversely proportional to the sum of the squares of the distance between the clusters, which is regarded as an error. In other words, the sum of the squares of the distance between the clusters must be low for two clusters to be joined together [4].

III. Preparation & Case Studies

Five load profile groups used by the Energy Market Operator of Turkey (EXIST) are commercial, industrial, residential, agricultural irrigation, and street lighting profile groups. Since the latter generally contains the same data every day, this group was not included in the study. Agricultural Irrigation profile group was not included in this study as well since the consumption values vary seasonally from region to region and therefore is not suitable for clustering. The following criteria are used regarding the data that was collected on the remaining 3 profile groups:

• 365 days of hourly continuous smart meter data is required because the analysis is conducted on a yearly basis,

• Only consumers with annual energy consumption of 50 MWh are included. This is a relatively big amount of consumption especially for residential consumers, but errors are presumed to be less for bigger consumptions.

Kilowatt-hour (kWh) is used as the hourly data consumption measurement unit of the meters used in the research. Since the meter consumption data is kept in 24-hour slices, the data repetition time, i.e. frequency, is 24 hours. The data to be compared for each electric meter is the 24-hour consumption data of the other meters corresponding to the same day.

The above mentioned real electricity consumption values are treated as time-series data and clustering studies have been conducted using the hierarchical clustering (HC) method. The success of clustering is measured from various angles using Euclidian Distance and DTW, which are the distance measurement types. Then, the profile groups on the EXIST Transparency Platform [6] and the cluster groups found using the HC method are investigated via an algorithm developed on MATLAB.

A. Data Collection

The data used in the research is collected from EXIST with its permission for R&D purposes only with no discrimination of private information. The data indeed received by EXIST from the distribution companies on a monthly and regular basis. Since settlement is done on an hourly basis in the Turkish electricity market, the smart meters from which data is collected hold electricity consumption data on an hourly basis for 365 days for a total of 8,760 singular electricity consumption data over one year. For this research, data is collected from 3 different DSOs in different geographical parts of Turkey. A total of 1,213 smart meters meets the criterion of having more than 50 MWh of annual consumption. However, only 957 of these meters have 365 days of continuous data. Therefore, 256 smart meters are eliminated from the initial dataset.

B. Data Cleansing & Filtering

The two ways that raw data can be cleansed are: (i) complete deletion of data thought to be incorrect or an outlier, (ii) replacing the incorrect data with the average of data on both sides (interpolation). We observed two different incorrect data in the electric consumption data collected from 957 meters. The first one is the missing data (or the data is

zero) for some hours, and the second one is the unexpected high consumption at some hours. Considering that there may be a power outage or high power consumption during these hours, these data haven't been cleansed. Since high consumption data plays an important role in electricity generation estimation, the outliers may be an important player in this estimation process. Thus, it can be said that no data cleansing was conducted in this research.

The data filtering step helps to group the collected data more easily (to include the groups of the focus of interest and to exclude the other groups from the research). Electricity consumption data can be collected on a seasonal basis as well as on a locational basis. Since the data used in this research is location-based electricity smart meter data from DSOs with known operational perimeters, no location-based filtering is required.

C. Data Normalization

Data normalization ensures that the data collected at different scales are meaningfully comparable and that the measurement unit can be omitted after data is converted to a certain range. One of the purposes of converting the hourly electricity consumption data held in kilowatt-hour into a range between 0 and 1 makes the comparison of the electric consumption values on different days and hours. The other purpose is to finalize the hourly settlement process using the load profile multipliers on an hourly basis. There are mainly two types of normalization [7].

The first kind of normalization is obtained by converting the data kept as a vector into another vector with a norm or length of 1, usually used in linear algebra [7]. The values in the new vector obtained are thus converted to a number between 0 and 1. The second kind of normalization is a process involving calculations of standard deviations of the vector elements called "Z-scores". The former method is adopted in this research.

D. Studies conducted

Electric consumption data from 957 smart meters have been processed using steps explained in Section II to obtain the representative load profiles (RLP). In the first phase of the study, all of the RLP data from 957 smart meters have been processed for the whole year via an algorithm developed on MATLAB. For this first phase, only Euclidean distance and average-linkage criteria are used since these distance and linkage criteria require less computational power and are faster. The purpose of this first phase is to check whether there is a clear pattern distinction between different RLP clusters. 5,10, 15, and 20 clusters are selected as the final number of clusters and the results are examined. Looking at the patterns of the resulting clusters, it is observed that there are no distinctive pattern similarities between the clusters from which different load profile groups can be identified. The Dynamic Time Bending (DTW) criterion, which requires high computing power, is not included in this first study.

Because electricity consumption data collected for this study is collected from different distribution companies in Turkey's three different regions, in the next phase of the research, the RLPs are evaluated within each region and load profile group. For example, Industrial Medium Voltage (MV) RLPs in DSO-2 are evaluated with each other and the algorithm developed on MATLAB is run to group the RLPs in 5 clusters and generate the relevant 24-hour graph. In this second phase of the study, only profile groups that had data coming from at least 100 smart meters are used for reliability. The distribution of profile groups and the number of meters in the group is shown in Table I.

TABLE I. NUMBER OF SMART METERS PER DSO AND PROFILE GROUP

Distribution Company ID	Profile Group	Number of Smart Meters
DSO - 2	Industrial MV	102
	Commercial LV	102
	Residential	119
DSO - 3	Commercial LV	139

Using the Euclidean distance method, average-link, and Ward's link criteria for linkage, cluster sets are found for 365 days and visually compared with the RLPs published on the web site of EXIST. The same study is carried out using DTW as the distance method. The number of days when more than fifty percent of the RLPs are collected in one cluster alone is counted are given in Table II.

Purpose of finding the number of days where more than fifty percent of RLPs are contained in a single cluster can be summarized as follows:

1. To see and compare the cluster patterns formed using the different linkage and distance criteria and to observe if there are any abnormalities in the patterns found,

2. Evaluate the methods used according to the goals of the study.

In particular, some of the objectives aimed by the latter item are: (i) finding outliers, specifically, regarding electric

TABLE II NUMBER OF DAYS PER YEAR WHERE MORE THAN FIFTY PERCENT OF RLPS ARE IN A SINGLE CLUSTER

Distribution Company ID	Profile Group	Number of Meters	Distance: DTW Linkage: Average No.of Days > 50%	Distance: DTW Linkage: Ward No.of Days > 50%	Distance: Euclidean Linkage: Average No.of Days > 50%	Distance: Euclidean Linkage: Ward No.of Days > 50%
DSO-2	Industrial MV	102	141	53	321	69
	Commercial LV	102	322	109	361	156
	Residential	119	365	200	365	186
DSO-3	Commercial LV	139	329	188	365	170

consumption, (ii) noticing clock drifts (if any), (iii) identify criteria that better analyze similarities by finding more uniformly distributed clusters, (iv) determining the speed and computing power requirements of the criteria used.

IV. Results & Analysis

There are many clustering methods and different variations of these methods in the field of data mining. Various studies have been conducted on clustering in the literature on electricity consumption and load profiles. Electricity consumption is a time series that repeats itself at a certain frequency, and it has been revealed in various studies that "Hierarchical Clustering" is one of the most suitable and well-known methods for clustering load profiles with similar patterns [4, 11, 12]. In this research, electric consumption data from 3 different regions in Turkey have been investigated and compared using the hierarchical clustering method. As distance criteria, Euclidian distance and DTW have been used; and for linkage criteria, average-link and Ward's links have been used. One of the important advantages of the hierarchical clustering method is that it is not mandatory to specify a certain number of clusters before the algorithm run. It is not practical for electricity distribution companies to determine an optimal or high number of clusters for load profiles. Therefore, in this study, the minimum number of clusters that are also encountered in most of the literature is chosen, which is 5. In Fig 2 to Fig. 5, graphical outputs for RLPs are given for a weekday on Jan. 2018. As mentioned previously, a preconfigured cluster number of 5 is chosen and output plots are drawn for the following combinations:

- Euclidian distance, average linkage,
- Euclidian distance, Ward's linkage,
- DTW distance, average linkage,
- DTW distance, Ward's linkage.

As shown in Fig. 2, Euclidian distance with average linkage column has a single cluster having more than 50 percent of all the RLPs for almost the entire year. This analysis can easily be seen by looking at the patterns and distribution of RLPs in a typical weekday on Jan. 2018 for DSO-2's industrial medium-voltage profile group in Fig. 2. In this figure, 78% of all the RLPs are collected in cluster number 4 and the remaining RLPs are distributed in the remaining four clusters. Three of these clusters can be classified as outliers with less than 3 percent of all the RLPs. Although one days' graphic is provided to illustrate the table's findings, similar results are acquired for the rest of the year as there are 320 more days in the year with similar figures for the same DSO and different profile groups as well as DSO-3 and Commercial LV profile group. Thus, Euclidean distance with average linkage criteria is a good combination to aggregate most RLPs in a single cluster and find outliers. This combination requires the least computation power and produces results significantly faster than the other combinations.

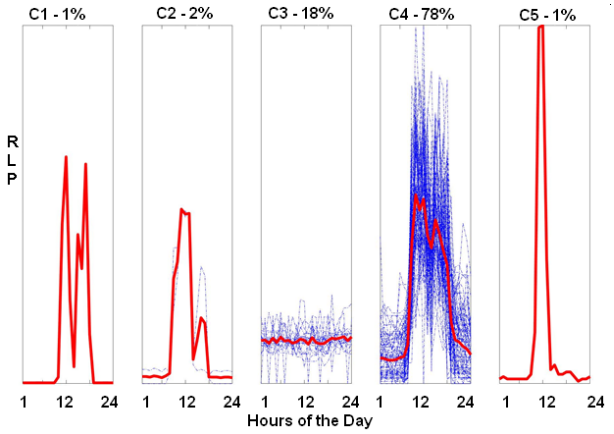

Fig. 2 A weekday cluster output for Euclidian distance and Average-linkage criteria for DSO-2 Industrial Medium Voltage

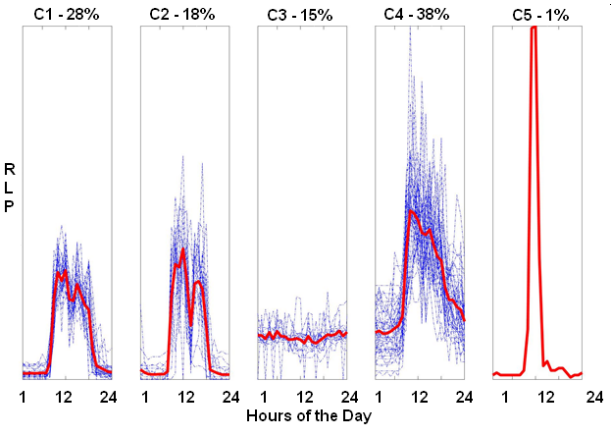

Fig. 3 A weekday cluster output for Euclidian distance and Ward's linkage criteria for DSO-2 Industrial Medium Voltage

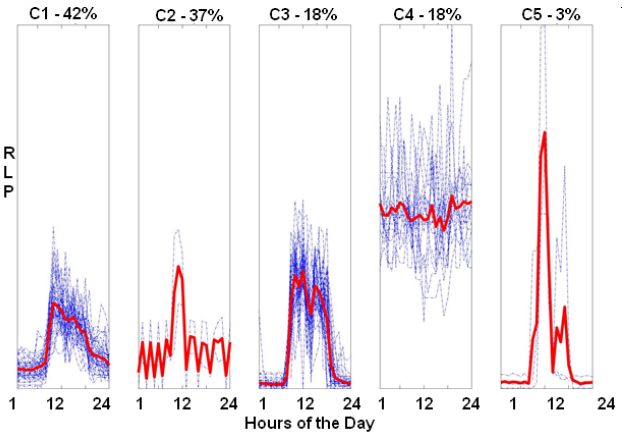

Fig. 4 A weekday cluster output for DTW distance and Average linkage criteria for DSO-2 Industrial Medium Voltage

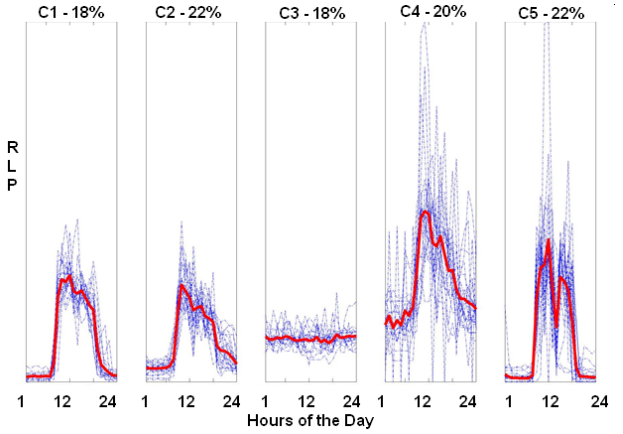

Fig. 5 A weekday cluster output for DTW distance and Ward's linkage criteria for DSO-2 Industrial Medium Voltage

Looking at Table II and the graphics (Fig. 3, Fig. 5) where Ward's linkage criterion is used, it is observed that there is a better distribution of RLPs between clusters. Also, the number of clusters with more than 50% of the RLP population is much less. The number of days per year where more than fifty percent of RLPs are in a single cluster is much less compared to the average linkage criterion, regardless of the distance metric. By looking at Table II, it can be said that the type of profile group also affects the output significantly. For example, industrial MV and residential profile groups are clustered significantly differently for Ward's linkage.

If the results were to be analyzed in terms of imbalance and settlement, it can be said that regarding imbalance, the study's findings are irrelevant. However, for settlement, although the study's output would have no impact on the overall settlement figure, there would be significant impact especially for consumers that don't have smart meters capable of holding and transmitting hourly consumptions, who are consuming relatively more electricity yet are profiled like the small consumers. The reason the study's findings would have no impact on imbalances is that the load profiles that are published on the market operator's transparency website are calculated with overall consumption whereas this study evaluates individual consumptions, regardless of consumption rates due to normalization. On the other hand, the settlement is a process between the market operator, distributors, traders, and consumers. Therefore, although the overall settlement figure would not change, the settlement between consumers and the distributors rely heavily on RLPs.

V. CONCLUSION

In this study, Turkish electricity consumption values were investigated by comparing Representative Load Profiles (RLP) s from different regions and profile groups using the hierarchical clustering method. Electric consumption data gathered at the Energy Market Operator of Turkey (EXIST) from 3 different electricity distribution companies were anonymized. Data from a total of 957 smart meters having continuous meter readings for the entire year and a total consumption above 50 MW per year were selected. No cleansing on the acquired data was done to find outlier data where available. The acquired consumption data was later normalized between 0 and 1 to find RLPs.

The initial phase of the study involved analyzing annual RLPs of all the data (collected from 957 smart meters per year) via a program developed on the MATLAB platform. Since the computational power necessary to evaluate the data for this initial phase was high, only Euclidian distance and average linkage criterion were used. 5, 10, 15, and 20 clusters were selected as the final number of clusters and the results were examined. Looking at the patterns of the resulting clusters, it was observed that there were no distinctive pattern similarities between the clusters from which different load profile groups can be identified.

In the second phase of the study, only profile groups that had data coming from more than 100 smart meters were used. Since it is not practical for electricity distribution companies to determine an optimal and high number of clusters for load profiles, instead of trying to find an optimal number of clusters, a cluster number of 5 was used in this study (which is usually the minimum number of clusters used in similar research). Euclidean distance with average linkage criterion was found to be a good combination to aggregate most RLPs in a single cluster and find outliers. It was observed that for Ward's linkage, there was a better distribution of RLPs between clusters. It was also found that the type of profile groups affected the cluster distribution as well. For example, industrial MV and residential profile groups were clustered significantly differently for Ward's linkage.

In summary, the hierarchical clustering method, using Euclidian and DTW criteria for distance and average and Ward's criteria for linkage, was a practical way to cluster and analyze RLPs. Average-linkage was better in finding well-populated clusters and outliers simultaneously. Whereas Ward's linkage criterion was better if the goal in clustering the RLPs was to find better-distributed clusters. It was also found that cluster distribution depended on the type of profile groups as well. Finally, although the study does not have any value-add for preventing imbalances in general, it may have a social benefit to consumers by introducing an adjustment process for settlement between consumers who are profiled and the relevant distributors.

ACKNOWLEDGMENT

This study is part of Murat Gunsay's Ph.D. dissertation. The authors would like to thank EXIST for sharing the dataset.

REFERENCES

[1] Mutanen, Antti. Customer Classification and Load Profiling Method for Distribution Systems. IEEE Transactions on Power Delivery, 26, 3, pp. 1755-1763, 2011.

[2] Wang, Xiaozhe, Miles, Kate Smith ve Hyndman, Rob J. Characteristic-Based Clustering for Time Series Data. Data Mining and Knowledge Discovery, pp. 335-364, 2006.

[3] Taylor, J. W. "Using Combined Forecasts with Changing Weights for Electricity Demand Profiling", The Journal of the Operational Research Society, s. 77-82, 2000.

[4] Rajabi, A. "A comparative study of clustering techniques for electrical load pattern segmentation", Journals & Books, 2020.

[5] Wang, Y. "Load Profiling and Its Application to Demand Response: A Review", Tsinghua Science and Technology, pp. 117-129, 2015.

[6] EXIST Transparency Platform. Available online: https://seffaflik.epias.com.tr/transparency/index.xhtml (accessed in 2020, June 6).

[7] Abdi, H. "Normalizing Data". Texas : The University of Texas at Dallas, 2010.

[8] Akperi, B., Matthews, P. "Analysis of clustering techniques on load profiles for electrical distribution", IEEE 2014 International Conference on Power System Technology, 2014.

[9] Attewell, P., Monaghan, D. B. and Kwong, D. "Data Mining for the Social Sciences: An Introduction", Data Mining for the Social Sciences. California : University of California Press, pp. 196-215, 2015.

[10] Everitt, B. "Cluster Analysis". Londra : Wiley, 2011.

[11] Chicco, G., Napoli, R. and Piglione, F. "Comparisons Among Clustering Techniques for Electricity Customer Classification". IEEE Transactions on Power Systems, pp. 933-940, 2006.

[12] Bidoki, S.M., Mahmoudi-Kohan, N., Gerami, S. and Bandar A. "Comparison of several clustering methods in the case of electrical load curves classification". IEEE 16th Electrical Power Distribution Conference, 2011.