A Three-stage Scheme for Consumers' Partitioning Using Hierarchical Clustering Algorithm

Antonia Nasiakou, Miltiadis Alamaniotis, and Lefteri H. Tsoukalas
School of Nuclear Engineering, Purdue University
Applied Intelligent Systems Labroratory
West Lafayette, IN, USA
{anasiako, malamani, tsoukala}@purdue.edu

Georgios Karagiannis
Department of Mathematical Sciences
Durham University
Durham, United Kingdom
georgios.karagiannis@durham.ac.uk

Abstract—The clustering of any type of consumers (residential, commercial, industrial) is of great importance in the operation of Smart Grids. In this paper, we propose a three-stage hierarchical scheme for residential consumers' partitioning using the Hierarchical clustering algorithm. The aim of this study is to cluster the consumers in well-separated and compact clusters using information from the near past (almost real time). The usage of electricity from a resident to another varies and this information can be used from the system operator for improving the efficiency of the distribution network. The first stage corresponds to the consumers' clustering of the distribution grid using data driven every three minutes (simulation time) from the meter of each residency. The procedure of the second stage takes part every a specific number of hours, called *h,* that is defined by the user. The average value of each of the *k**20 clusters formed the last *h* hours is used as input for the hierarchical algorithm. In the third stage and in the end of each *h* hours, the average value of the data in each cluster is calculated and each consumer is reassigned to the cluster where the a distance metric is minimized. The results of the second stage provide deeper information about the load patterns existing each hour in the distribution grid. This information can be used from suppliers to design the energy tariffs for suiting better to the consumers' needs. This approach is tested using the IEEE-13 test feeder. The data are driven from 56 residencies.

Keywords—*consumers; partitioning; hierarchical clustering; real-time; smart grids*

I. INTRODUCTION

The last decade, there has been paying more attention in the Smart Grid and its applications. In comparison with the present grids, smart grid can efficiently transmit the electricity, restore the electricity after power disturbances (e.g., blackout) and offer a balance between the supply and demand through the participation of the consumers in the operation of the grid. The basic component of the smart grid is the smart meters. A huge amount of data can be driven in real time using the smart meters. In this study, we analyze the data from the smart meters located at distribution grid end points (e.g., residential). This data can be utilized by machine learning techniques. More specifically, clustering algorithms are utilized to partition similar consumers considering their consumption habits.

The clustering of the consumers has a plethora of applications. First of all, it facilitates the data mining and analysis of the data obtained from the smart meters [1] and the decision making method [2]. It also eases the assignment of the energy tariffs from the suppliers to better suit the energy behavior of each consumer of a distribution grid. Furthermore, the clustering results can be used for load forecasting, short-term or medium-term and for Demand Response (DR) programs [3].

In this study, we propose a three-stage hierarchical scheme as its architecture is depicted in Fig. 1.

1. In the first stage, the hierarchical clustering algorithm is performed giving groups of consumers using load data driven every 3-minutes.
2. In the second stage, the average value of each of the *k**20 clusters formed the last *h* hours is used as input for the hierarchical algorithm. *k* is the number of clusters.
3. In the last stage, a distance metric of the load value of each consumer with the average value of each of the *k* clusters is calculated. The consumer is reassigned to the cluster with the minimum distance.

Our main contribution is the proposal of at almost real time clustering of the consumers using: 1) the hierarchical clustering algorithm and 2) a distance metric to reassign the consumers to the most appropriate cluster.

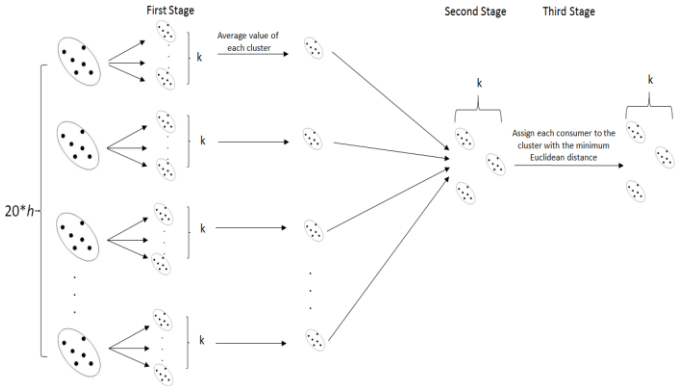

Figure 1 Architecture of the proposed hierarchical scheme, where h is the hours the proposed scheme is applied

Comparing to the other efforts where the consumers clustering is performed using historical data, in this study every 3-minutes data are used to eventually cluster the consumers into groups using data from the last twenty 3-minutes time intervals for the *h* hours. It is worth mentioning here that we compare the clustering results of the proposed scheme with the results obtained when the three-stage scheme is not used to prove that using more almost real time information the resultant clusters are better using validity indicators.

The remaining of the paper is organized as follows. In Section II, an overview of the related work consumers' clustering is presented. In Section III, the proposed scheme is described in detail and the results through the experiments are presented in Section IV. The concluding remarks are summarized in Section V.

II. Basic Concepts and Related work

A. State of the art

Recently, many methodologies and approaches have been studied in the literature to group the consumers of a power grid using clustering algorithms, including k-means [4], self-organized maps [5], hierarchical [6], and fuzzy c-means [7]. For assessing the validity of the clustering results of each algorithm, various validity indicators have been developed. The most commonly used indicators are the clustering dispersion indicator, the mean index adequacy, the similarity matrix indicator, the Danies-Bouldin, the Dunn index, the scatter index and the mean square error. In addition to the clustering algorithm, data mining techniques, wavelet packet transformation are also utilized for grouping of the consumers.

More precisely, in [3] and [8] the authors propose a two stage pattern clustering of electricity customers. Both studies in the first stage, cluster the daily load patterns of each consumer. In the second stage, the clustering of the consumers is performed by using as input for the clustering algorithm a representative load pattern from each consumer. The main difference between those two approaches is that he authors in [3] use the technique of the Fast Wavelet Transformation to reduce the dimensionality of the data used for the clustering, and the g-means algorithm instead of the k-means to adaptively select the number of clusters that really exist. Their results prove that the proposed method provides a more stable system. On the same hand the authors in [8] use various validity measures to prove the superiority of their proposed methodology over the other clustering algorithms.

In [9] the authors try to group a set of 234 non-residential consumers in Italy. In this study, various validity measurements are used to compare several clustering algorithms. A similar effort has been conducted in the study presented in [10] where a clustering procedure takes place in a building in a Greece's university campus.

The authors in [11] use means of dynamic clustering to cluster and analyze the load patterns of a set of Spanish residential consumers. More specifically, the authors use the k-means algorithm to cluster three different types of consumers; the first one is characterized by three peaks during
a day, the second one is characterized by a quasi-flat load pattern during a day and the last one contains consumers that use more electricity during the night.

Moreover, some other efforts in the literature use the clustering techniques for prediction purposes. More specifically, in [12] the authors prove that the clustering of the consumers may give more accurate load forecasts. In the same way in [13] the authors use a sequential cluster weighted modeling (SCWM) to improve load predictions by recognizing specific load patterns.

B. Hierarchical Clustering Algorithm

The hierarchical clustering algorithm clusters the data at the same time creating in that way a cluster tree. Fig. 2 shows such a tree. The tree is actually a hierarchy where the clusters are created by joining clusters at each level. The advantage of hierarchical clustering algorithm over the other clustering algorithms is that the data is kept without any changes. The hierarchical clustering runs as follows:

1. In the beginning, each data point corresponds to a cluster
2. The most similar pair of clusters is merged into one cluster based on the similarity criterion
3. Find the similarities between the new cluster and the old ones
4. Repeat step 2 and 3 until the cutting criterion is met

The grouping of the data points is based on linkage criterion such as average distance, shortage distance, centroid distance and Ward distance.

C. Evaluation Metrics

Various evaluation metrics, such as the cluster dispersion indicator (CDI), Davies-Doulbin index (DBI), the similarity matrix indicator (SMI) and the Dunn index (DI) have been widely utilized for evaluating the clustering of load patterns.

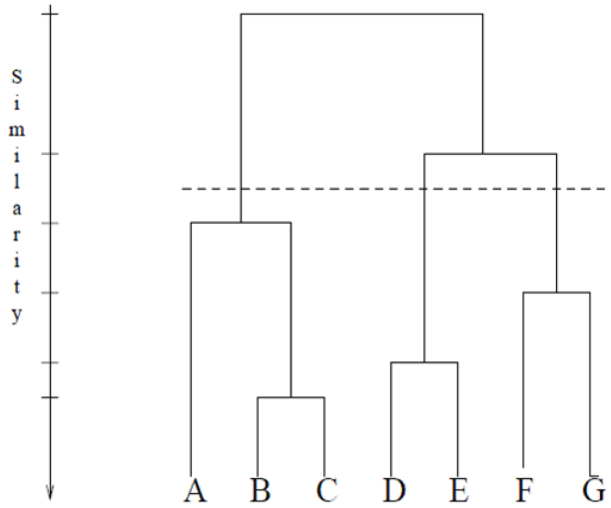

Figure 2 A tree obtained using Agglomerative Hierarchical Algorithm [14]

In this subsection these metrics and the distances that are used from them are described [3].

1. Distance of the items in a cluster

$$d(C_j) = \sqrt{\frac{1}{2|C_j|}\sum_{x_i \in C_j} d(x_i, C_j)^2} \qquad (1)$$

2. Distance of a vector and a cluster

$$d(x_j, C_j) = \sqrt{\frac{1}{|C_j|}\sum_{x_i \in C_j} d(x_i, x_k)^2} \qquad (2)$$

3. Distance between two vectors

$$d(x_i, x_j) = \sqrt{\frac{1}{n}\sum_{k=1}^{n} d(x_{i,k} - x_{j,k})^2} \qquad (3)$$

The DBI gives the average of each cluster's similarity measurement with its most similar cluster. A distance metric is used to calculate the distance between pairs of the cluster centers and these distances are used to calculate the SMI. In (4) and (5) the mathematical formula for the DBI and SMI are presented respectively.

In the following equations, the number of cluster centers indicated as R and the number of clusters as K.

$$DBI = \frac{1}{K}\sum_{k=1}^{K} max\left\{\frac{d(x_i, C_k) + d(x_j, C_k)}{d(R)}\right\} \qquad (4)$$

$$SMI = \max_{\substack{i>j \\ i,j \in \{1,\ldots,K\}}} \left\{\left(1 - \frac{1}{\ln d(\mu_i, \mu_j)}\right)\right\} \qquad (5)$$

The CDI is defined as follows:

$$CDI = \frac{1}{d(R)}\sqrt{\frac{1}{K}\sum_{k=1}^{K} d(C_k)^2} \qquad (6)$$

The DI provides clusters that are compact and well separated. The value of the DI can be calculated using [gmail] the formula below:

$$DI = \min_{1 \le i \le j \le K} \frac{d(C_i, C_j)}{\max_{1 \le k \le K} d(C_k)} \qquad (7)$$

For all the above validity measures, the compactness the well-separation of the clusters is their priority. It is worth mentioning in that point that in this study for evaluating the clustering results of the proposed scheme the DBI is used as its implementation is available in MatLAB.

D. *GridLAB-D Simulation Platform*

In this study, the GridLAB-D simulation platform is used to obtain the data for the clustering procedure. GridLAB-D is an open source platform for simulating the operation of the distribution grid. It is characterized by its high performance and the capability of allowing the users to extract, modify or add their own modules. The data obtained from each simulation are time series data which helps the user to read them. Actually, a simulation engine based on the GridLAB-D configuration is developed.

III. THREE-STAGE HIERARCHICAL SCHEME

The benefits of our proposed scheme are as follows

- The three-stage hierarchical scheme clusters the consumers in well-separated and compact clusters using information from the near past (almost real time)
- The resultant partitions include consumers with similar energy consumption behavior due to the fact that the minimum distance is utilized for the final assignment to the resultant clusters.
- We apply the hierarchical algorithm for the partitioning using as input the total demanded load of each consumer

A. *Every 3-minutes clustering procedure-Stage 1*

This stage comprises two stages:

A. clustering of the residential consumers of the distribution grid
B. calculation of the average value of all the values per cluster.

More precisely, the measurement data are obtained through the meters placed to residential consumers and they correspond to time series data which represent the total demanded load of the under consideration set of consumers. The measurements for this stage are performed on a 3 minute basis. The hierarchical algorithm uses these measurements as input to cluster the consumers in a 3 minute basis. For the time interval of the *h* hours, resulting in 20**h* sets of *k* clusters, the average value of the items belong to each cluster formed for the particular hours is calculated. So, in the end of this stage, the information that is going to be used for the next stage is 20**k***h* item values.

B. *Every h-hours clustering procedure-Stage 2*

In this stage before conveying through the next *h* hours of the simulation time, the hierarchical algorithm is applied. This time, the hierarchical algorithm uses as input the output of the previous stage. Therefore, when the hierarchical algorithm gives the clustering results, the output of this stage can be used for the procedure of the third stage. The output of this stage is *k* clusters which provide information about the average demanded load of the consumers the last *h* hours of the simulation.

C. *Reassignment of consumers in the new clusters-Stage 3*

In this stage, the minimum distance is used for the assignment of the consumers in the final clusters. The data points in this study are one dimensional. In (7), the mathematical formula of the minimum distance for 1-D data points is presented.

$$d(x_i, C_j) = x_i - C_j \qquad (7)$$

where x_i is the ith consumer of the distribution grid with the i goes from 1 to N (the number of consumers) and Cj is the jth cluster with the j from 1 to *k* (the number of clusters). In particular, the algorithm which assigns the consumers to the final clusters for the under consideration hour works as described below in Algorithm A where *k* is the number of clusters and N the number of consumers.

Algorithm A: find the minimum distance **Input**: x_i's value and average value of all the Cj **Output**: reassignment of each x_i to the cluster Cj with which x_i has the smallest distance
1. for each i=1, …, N 2. for each j=1, …, k a. Find $d_j = d(x_i, C_j)$ b. Take the minimum d_j with the j to be the id of the cluster c. Assign the x_i to the j cluster

IV. EXPERIMENTAL RESULTS

For demonstrating and evaluating the proposed three-stage scheme, we use a dataset covering one day on 3 minute resolution load measurements at the IEEE-13 test feeder (see Fig. 3). We configure the load that the feeder can afford, so the number of residencies-consumers with which the IEEE-13 test feeder is equipped is 56 low voltage ones. The 23 residencies out of the 56 are equipped with an hvac (heating-ventilation and air conditioning) system, a water heater and the refrigerator. The remaining 23 residencies are equipped only with refrigerators. We configure in such way the consumption of the residencies to show how the proposed scheme works.

For comparison purposes and for assessing the validity of the proposed scheme we consider two cases. In the first case, called Case **I**, the proposed three-stage scheme is applied and the final clusters are the ones obtained every *h* hours in a 24 hour resolution. In the second case, called Case **II,** the partitioning of the distribution grid takes place every *h* hours without usong the proposed scheme. So, the clustering results for both cases are compared using the Davies-Doulbin clustering measure. It is worth mentioning in that point, that in the Case **II,** the proposed scheme is not applied. In particular, the consumers are grouped into clusters using the hierarchical algorithm which takes as input the load measurements of the particular hour.

In the experiments presented in this study the clustering of the second stage has taken place every 6 hours and the number of clusters for each stage of the proposed scheme is 2. We have selected just two clusters due to the low number of consumers and for easier presentation of the results. So, the number of inputs of the hierarchical algorithm at the second stage is 20**2*6. In Fig. 4 and 5 the resultant clusters for the Case **I** are presented. These results correspond to 6:00 am. The results of Case **II** for the same time interval are presented in Fig. 6 and 7. It is obvious that when the proposed scheme is utilized to cluster the consumers for a specific hour, the consumers that belong to each cluster have almost the same energy consumption requirements. More specifically, the peaks of each consumer may indicate the use of the hvac system, so in the same clusters belong consumers that use the

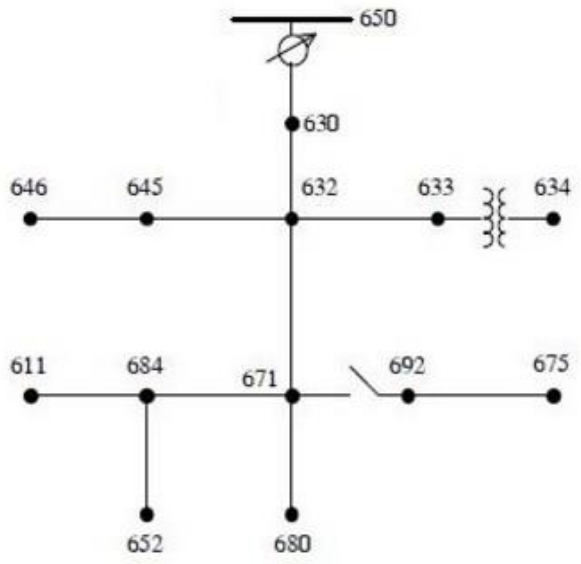

Figure 3 IEEE-13 test feeder architecture [15]

hvac system same number of times during a specific time interval.

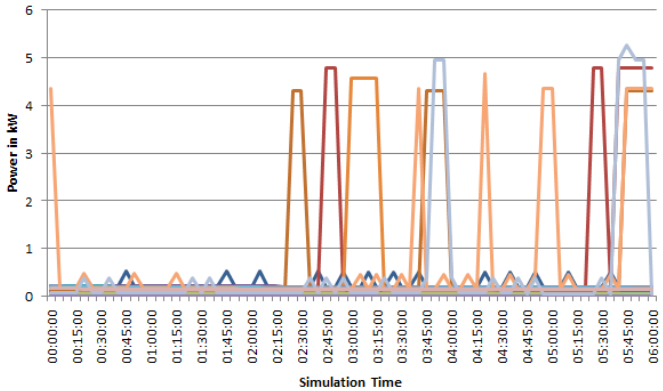

Figure 4 Items belonging to cluster A of the Case I

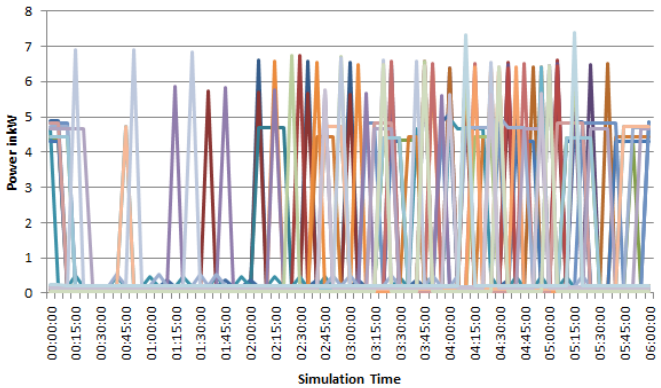

Figure 5 Items belonging to cluster B of the Case I

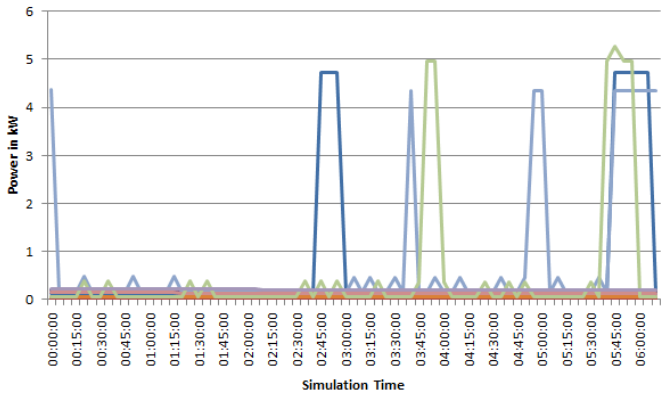

Figure 6 Items belonging to cluster A of the Case II

It is worth to be mentioned at that point that due to the fact that it is not clear from the figures because of the volume of the data, the number of consumers of Case I that belong to cluster A and cluster B is 20 and 36 respectively. On the other hand, the number of consumers of Case II is 11 and 45 for the cluster A and cluster B respectively. Therefore, the consumers are evenly distributed in both the clusters for the Case I. Additionally, when the clustering procedure took place at 12:00pm the number of consumers of Case I and II is 35 and 7 for the cluster A and 21 and 49 for the cluster B respectively. So, we can tell that using the proposed scheme more consumers that have similar energy consumption habits are assigned to the same cluster.

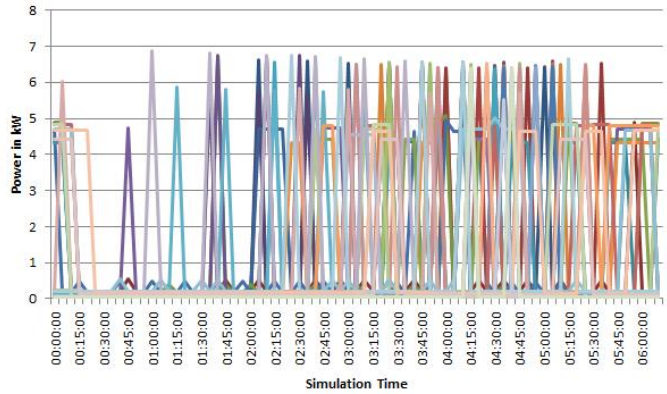

Figure 7 Items belonging to cluster B of the Case II

TABLE I Validation indices results

	DBI (Case I)	DBI Case II)
6:00am	1.41	1.47
12:00pm	1.36	10.63
6:00pm	1.59	2.58
12:00am	1.24	1.46

We model the distribution grid in such a way to consider if the consumers with the same energy consumption requirements assign to the same cluster. In particular, by using the information obtained every 3 minutes for the interval of 6 hours instead of using only the information which is available the specific hour provides better clustering results as we can see. More specifically, in the Case I the clusters have approximately the same number of consumers. However, the ideal will be occurred if each cluster has 23 consumers with exactly the same energy consumption characteristics. This is something that is in our future plans.

Moreover, as it can be observed from the results of Table I, it is clear that the proposed three-stage scheme provides better clustering results comparing the DBI for both cases. The lower the value, the better the clustering results. Therefore, using information for the last 6 hours instead of using the current information about the load of each consumer helps the clustering of the consumers in groups with almost the same habits.

In the next lines, we will present some more experimental results, by changing the time interval that the final clustering takes place. In these results, the final clustering took place every 1 hour and in the Fig. 8, 9, 10 and 11 we can see the results for the Cluster A and B of Case I and II respectively at 1:00pm. We choose the time interval to be too short to prove that the proposed scheme works even when the information obtained using a short time interval.

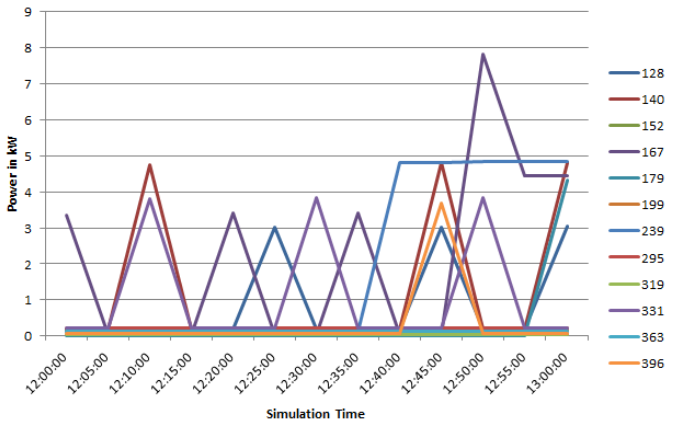

Figure 8 Items belonging to cluster A of the Case I

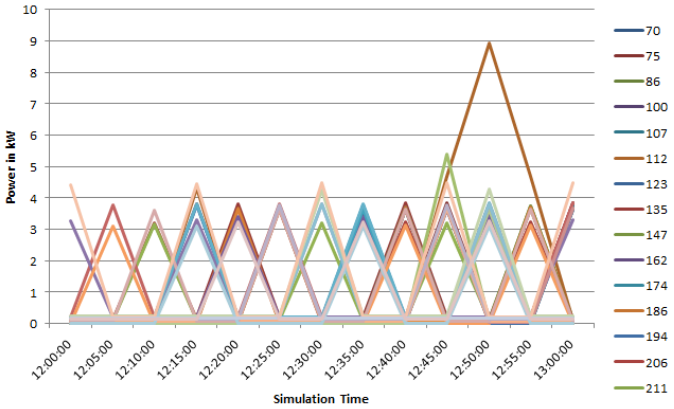

Figure 9 Items belonging to cluster B of the Case I

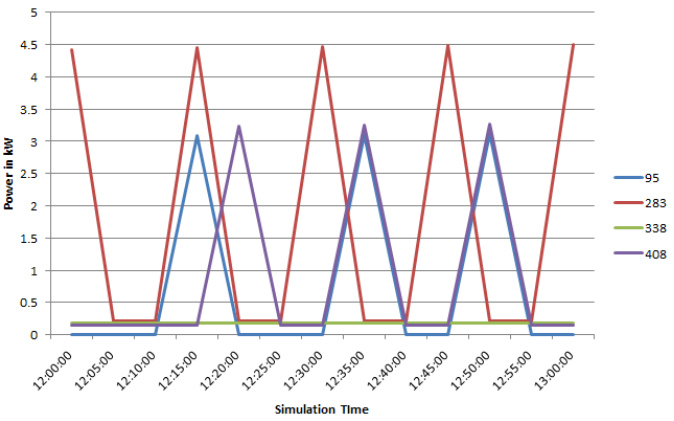

Figure 10 Items belonging to cluster A of the Case II

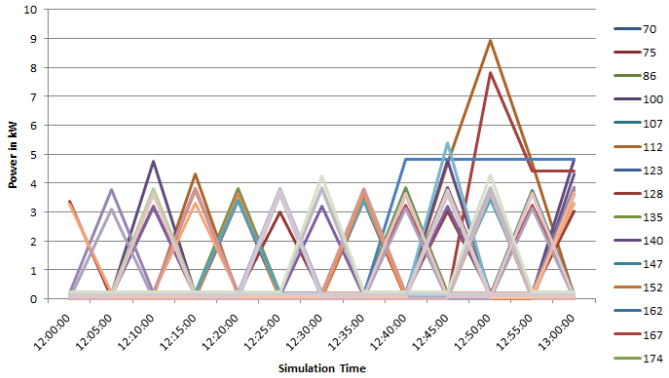

Figure 11 Items belonging to cluster B of the Case II

However, it is observable, that the consumers are not distributed as well as they are distributed in the case where the time interval is 6 hours. In particular the number of consumers is 44 and 52 for the cluster A of the Case I and II respectively. The difference is just 8 consumers that are assigned to the cluster B of the case I. For example, at 12:00pm the cluster A has 35 and 7 consumers in the case I and II, respectively; a difference that is notable. This is a side effect of the fact that the information in this case is less than in the case where the time interval is 6 hours.

V. CONCLUSIONS

In this study, a three stage scheme for clustering of residential consumers is introduced and described in detail. We apply the proposed scheme in a set of load data obtained through a simulation conducted using the GridLAB-D simulation platform. The main contributions of the proposed scheme are as follows

1. Use of data from the near past to group the consumers
2. The number of consumers in each cluster is evenly distributed when the proposed scheme is applied

The results of using the proposed scheme prove that the clustering of the consumers can be achieved even without using any historical data only data from the near past. It can be noticed that the clustering of consumers can be conducted almost real time by using data only from the near past and not historical data. To assess the validity of the proposed scheme the Danies- Doulbin and Dunn indices are used to measure the compacteness and the degree of the good separation of the clusters. For all the results presented in this study the above validity measures have better values; the lower the values, the better the clustering results.

REFERENCES

[1] G. Chicco, “Overview and performance assessment of the clustering methods for electrical load pattern grouping,” Energy, vol. 42, no. 1, pp. 68–80, Jun. 2012.

[2] G. Chicco, R. Napoli, P. Postolache, M. Scutariu, and C. Toader, “Customer characterization for improving the tariff offer,” IEEE Trans. Power Syst., vol. 18, no. 1, pp. 381–387, Feb. 2003.

[3] K. Mets, F. Depuydt, and C. Develder, “Two-Stage Load Pattern Clustering Using Fast Wavelet Transformation,” IEEE Trans Smart Grid, vol. 7, no. 5, pp. 2250-2259, Sept. 2016.

[4] A. Nasiakou, M. Alamaniotis, and L.H. Tsoukalas, “Power Distribution Network Partitioning in Big Data Enviroment using k-means and Fuzzy logic,” Mediterranean Conference on Power Generation, Transmission, Distribution and Energy Conversion, pp. 89-97, Nov. 2016

[5] S. V. Verdu, M. O. Garcia, C. Senabre, A. G. Marin, and F. J. G. Franco, “Classification, filtering, and identification of electrical customer load patterns through the use of Self-Organizing Maps,” IEEE Trans. Power Syst., vol. 21, no. 4, pp. 1672–1682, Nov. 2006.

[6] S. M. Bidoki, N. Mahmoudi-Kohan, M. H. Sadreddini, M. Zolghadri Jahromi, and M. P. Moghaddam, “Evaluating Different Clustering Techniques for Electricity Customer Classification,” Transmission and Distribution Conference and Exposition, 2010 IEEE PES,pp. 1-5, Apr. 2010.

[7] N. Anuar, and Z. Zakaria, “Determination of fuzziness parameter in load profiling via Fuzzy C-Means,”Control and System Graduate Research Colloquium (ICSGRC), pp. 139 – 142, 2011.

[8] G. J. Tsekouras, N. D. Hatziargyriou, and E. N. Dialynas, “Two-Stage Pattern Recognition of Load Curves for Classification of Electricity Customers,” IEEE Trans. Power Syst., vol. 22, no. 3, pp. 1120-1128, Aug 2007.

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

[10] I.P. Panapakidis, T.A. Papadopoulos, G.C. Christoforidis and G.K. Papagiannis, “Analysis of the Electricity Demand Patterns of a Building in a University Campus,” IEEE 12th International Conference on Environment and Electrical Engineering, 2013.

[11] I. Ben´ıtez, A. Quijano, J.-L. D´ıez and I. Delgado, “Dynamic clustering segmentation applied to load profiles of energy consumption from Spanish customers”, Electrical Power and Energy Systems, vol. 55, pp. 437-448, 2014.

[12] S. Humeau, T.K. Wijaya, M. Vasirani and K. Aberer, Electricity Load Forecasting for Residential Customers: Exploiting Aggregation and Correlation between Households,” 3rd IFIP Conference on Sustainable Internet and ICT for Sustainability, Palermo, Italy, 2013.

[13] T. Zhu, S.R. Shaw and S.B. Leeb, “Electric load transient recognition with a cluster weighted modeling method,” IEEE Transactions on Smart Grid, vol. 4, no. 4, pp. 2182-2190, 2013

[14] A. K. Jain, M. N. Murty, P. J. Flynn, "Data Clustering: A Review". ACM Computing Surveys (CSUR), vol. 31, no. 3, pp. 264-323, 1999.

[15] R. Fainti, M. Alamaniotis, and L.H. Tsoukalas, “Three-phase congestion prediction utilizing artificial neural networks,” Information, Intelligence, Systems & Applications (IISA), 2016 7th International Conference on, pp.1-6,Jul.2016.