Dynamic Clustering Segmentation Applied to Load Profiles of Energy Consumption from Spanish Customers

Ignacio Benítez[a], Alfredo Quijano[a], José-Luis Díez[b], Ignacio Delgado[a]

[a]*Energy Technological Institute*
Avda. Juan de la Cierva 24
46980 Paterna, Spain
[b]*Department of Systems Engineering and Control*
Universitat Politècnica de València
Camino de Vera, 14
46022 Valencia, Spain

Abstract

The following article describes the work of dynamic segmentation of daily load profiles throughout years 2008 and 2009, of a representative sample of Spanish residential customers. The technique applied is classification of the energy consumption time series of load profiles by means of dynamic clustering algorithms. The techniques used and analysis performed prove adequate as a fast tool to classify clients according to their energy consumption patterns, as well as to evaluate their overall energy consumption trends at a glance. The segmentation of the energy consumption load profiles is performed, and the results are analyzed and discussed.

Keywords:
dynamic clustering, data mining, demand side management, load profiles

1. Introduction

The new specifications that arise in the energy market make necessary an approach to an effective measurement and management of the end-user energy consumption and trends, not only concerning the traditionally supervised large consumption customer, but also the medium and high energy consumption residential user, whose consumption profile depicts unbalanced patterns of peaks of energy consumption, and valley or peak–off regions where the energy demand remains unsolicited.

The Demand Side Management (DSM) tools allow a more effective interaction of energy production and consumption profiles, therefore providing the end-user a valuable interface to achieve different levels of energy management. A definition of DSM or Demand Response (DR) can be taken, for instance, from the U.S. Department of Energy technical report: "Changes in electric usage by end-use customers from their normal consumption patterns in response to changes in the price of electricity over time, or to incentive payments designed to induce lower electricity use at times of high wholesale market prices or when system reliability is jeopardized [1]." DSM addresses management by financial incentives and awareness, whereas DR is related to the active management of loads in households and appliances.

As stated in different experiences [2] DSM tools are a valid strategy to contribute to a more adequate use and management of the energy, in one of the following objectives: peak shaving at specific hours through the day, lowering energy demand from the daily load profile, lamination of the demand to prevent a steep slope in energy consumption profile, or shifting the use of energy in residential customers from the peak to the valley hours of the day.

In this work, an analysis of the load profiles of a representative sample of Spanish residential users is applied, by means of dynamic clustering [3][4][5]. Dynamic clustering algorithms perform segmentation of time-series data in groups or clusters by similarity. These algorithms belong to a wide set of algorithms and techniques for data analysis, grouped and classified under the term of *data mining* [6][7][8].

A dynamic clustering algorithm is applied on a database of daily load profiles from 759 clients during years 2008 and 2009. The objective of this analysis is twofold: on one hand, to classify the Spanish residential users by their dynamic daily load profile, evaluating the influence of the changes in normative and laws in the Spanish energy market that were produced in years 2008 and 2009. On the other hand, this work wants to present the possibilities of the dynamic clustering analysis, which can be a very useful tool that can be used by experts to classify groups of clients at a glance, detecting abnormalities in the energy load profile, evaluating the trends, and selecting target groups for DSM actions.

1.1. Data mining and KDD

The term “data mining” can be found in the literature under a number of different definitions, all of them pointing to the same target, which is the analysis and extraction of useful information from large sets of data. The following description can be fitted to this objective: *“Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner [7]”.*

Data mining techniques have been described as the intermediate step within a bigger process, called *knowledge discovery in databases* or KDD, described as *“the nontrivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data [5]”.* This process covers the whole procedure of a data set analysis, in the following steps:

1. *Data warehousing*: the first step of the analysis, comprising all the techniques and procedures to conveniently process erroneous or missing values, and rearrange data in order to be analyzed.
2. *Data mining*: the intermediate step, has as objective to extract useful relations or information from the data.
3. *Interpretation of results*:As a final step, the obtained results from the data mining procedures are analyzed, usually by an expert, and conclusions are generated, that will help to fulfill the initial objectives for which the data was registered.

Although descriptions may not agree in some cases, commonly three different groups of data mining techniques are found in the literature [7]: classification, prediction and time series analysis. Classification techniques obtain models that distinguish data classes from a set of data. This can be a supervised learning process, where the model is compared and adjusted based on a reference model, or a non-supervised learning one, where the data are classified without reference models. Prediction models have as objective the production of models able to simulate and predict future outcomes of a set of variables, based on the values of other variables from the data set. Time series analysis apply classification and prediction techniques to evaluate the trends and evolution of time series or sequential data. Figure 1 depicts the areas and main techniques applied in the data mining process.

Concerning the analysis of load curves of energy consumption, a number of pattern recognition and classification studies have been done. In [9] the

Figure 1: Data mining techniques.

number of classes to be characterized is obtained from the tariff structure, which is directly related to the voltage consumption level and the usage made (residential, industrial, commercial) by the customers. From this information, a stratified random sampling is applied, choosing carefully the sample sizes to match a specific confidence level and standard deviation. In [10] this work is extended to other regions or districts, thus adding the geographical dimension to the study, allowing to infer the power system demand by aggregating typical patterns from customers, tailored for each district.

In [11][12][13][14] studies of classification techniques have been developed, including clustering methods and Self Organizing Maps (SOM) [15], in time domain and frequency domain. Identification of the loading conditions (e.g., season influence and a partition of days in working and non-working days) is suggested in this work, as a pre-treatment stage. This partition is considered in the present paper. Other works [16][17] also perform a comparative method of different classification techniques, validating the results with adequacy measures and indicators for the resulting partitions, such as the Davies - Bouldin or DB index [18].

Non-supervised and supervised classification methods have been developed in other works [19][20] to first obtain the different classes or prototypes of load profiles by a combination of SOM, K-means clustering [21][22], as the basis for a classifier of load profiles from electricity customers, based on decision trees or Artificial Neural Networks (ANN) [23]. In [24] the classification of loads is used to determine the economic activity, based on probability neural networks.

In the present work, a dynamic clustering algorithm has been applied on the data set of daily load profiles, therefore the objective is to classify and find a number of representative patterns in time series data. The dynamic clustering techniques are described next.

1.2. Dynamic clustering techniques

A complete definition of clustering can be found in the book by Oliveira and Pedrycz [25], where clustering is defined as the following:

Figure 2: Example of two clusters (A and B) and their respective centroids, (x_1, y_1) and (x_2, y_2), on a 2D data set.

> “Clustering is an unsupervised learning task that aims at decomposing a given set of *objects* into subgroups or *clusters* based on similarity. The goal is to divide the data set in such a way that objects (or example cases) belonging to the same cluster are as similar as possible, whereas objects belonging to different clusters are as dissimilar as possible [...]. Cluster analysis is primarily a tool for discovering previously hidden structure in a set of unordered objects [...]”

Clusters or classes are groups of objects, formed under some specified similarity criterion. Each cluster has usually an object that is representative or prototype of the objects pertaining to the cluster; this object is called the centroid. Figure 2 displays an example of clustering in a set of $2D$ objects. The centroids usually coincide with the center of gravity of the cluster, however this may vary, according to the algorithms used and the data types being classified.

Clustering and data mining techniques are usually performed on static data, bounded within a specific range of time when the data were acquired, and therefore valid for that specific range. A sample time consistency is expected, meaning that all the data being analyzed should share a predefined value of acquisition time or a fixed time range.

Dynamic data mining techniques, on the other side, deal with the analysis of dynamic, time-dependent changing data. These data can be differentiated, according to their nature, purpose and mining objectives, in three types [6]:

- **Data Streams**. This definition typically refers to massive amounts of multidimensional data, being continuously obtained from any source, in chronological order. Data mining techniques for data streams require an online processing, due to the huge amount of data and the complexity of storing it all and processing it off-line.

- **Time-series database**. These databases comprise sequences of data with time stamp, which have been obtained over repeated measurements of time. A typical example of time-series data are stock market prices.

- **Sequence database**. A sequence database is obtained from sequence measurements, which can be related to time or to other variables.

In this work, a dynamic clustering algorithm is applied off-line to a database of time series data of energy consumption load profiles. The dynamic data is formed by objects with dynamic features, represented by feature trajectories in time. The static data, on the contrary, is represented by objects with a features vector of values captured at a given time. The dynamic clustering algorithm performs segmentation, therefore, on a dynamic data set, augmenting the possibilities of clustering beyond a static "picture" of clusters at a given time or period. Table 1, extracted from Weber [25], classifies the cluster analysis in four types or categories, according to the dynamic nature of the data and the clusters:

1. The data, although is time series, is treated as static and the clustering process is also static. This is the case when the static clustering is applied, and corresponds to the example seen in Figure 2.
2. The data, although is time series, is treated as static. The number of clusters, however, is not fixed, and may vary at each new computation. This is the case when the database is partitioned in batches and processed sequentially. For each batch or sequence, the number of clusters may vary, indicating a variation in the data values or trends. This clustering process implies, in case a relationship among clusters is to be established through time, the need of a pattern matching or recognition system, able to associate the new clusters with the ones from previous steps[3]. In this system, issues such as clusters formation, collapse, split or fusion must be considered.
3. The data is treated as dynamic, evolving through time, therefore the time series become trajectories of the different data features or dimensions through time. The number of clusters is fixed. The objects are clustered taking into account the evolution or trajectories of the object features and, therefore, the resulting centroids or patterns are defined by feature trajectories that are representative of the data evolution in the different resulting clusters. The present paper approaches this type of dynamic clustering analysis.

Table 1: Types of clustering according to dynamic nature of data and classes

4. The data is also treated as dynamic, as in type 3, becoming feature trajectories that evolve through time, and the number of clusters varies dynamically at each iteration. Clusters and patterns can, therefore, as in type 2, merge or split.

Type 1 in Table 1 represents static clustering. Types 2 to 4 are considered dynamic clustering, due to the dynamic nature of data, classes, or both. For this work, the number of classes is fixed (static), addressing therefore a dynamic clustering analysis of the type 3 in the Table. Ongoing research is being done on dynamic clusters, updated at each iteration (type 4 in the Table).

1.3. Structure of this document

The present article is divided in the following sections:

- Material and methods, where the data used and the techniques applied are explained.
- Dynamic clustering segmentation, where the dynamic clustering process is described and the results displayed.
- Results and discussion, where the results obtained are analyzed and discussed.
- Conclusions, where the main outcomes from this work are gathered and presented, along with future works that could refine the analysis and provide an added value to this tool.

These sections are detailed next. The document is also completed with the inclusion of acknowledgements and references.

2. Material and methods

Following, the two components of this analysis will be described. These are: the data which has been used (format, origin) and the dynamic clustering algorithm that has been applied on the data.

2.1. Description of the database used from Spanish residential users

The database comprises hourly energy consumptions taken from smart meters of a number of selected customers through Spain, with the objective to become a representative sample of the entire population of residential energy consumers. Following, a brief information from the sample and the database is provided.

2.1.1. Sample of residential users

The sample of users represents typical residential load curves from Spanish consumers, in the three representative groups of customers of residential energy consumption: the so called "type a", "type b" and "type c" clients. These categories have been defined in the Spanish legislation, and are divided according to the type of contracted tariff and expected load profile:

- The "type a" clients are those who have contracted 2.0 A or 2.1 A tariffs. These are low voltage (< 1000 V) tariffs with no time-of-use (TOU) pricing. The 2.0 A tariff is for clients with a power consumption below or equal to 10 kW, and the 2.1 A tariff is for clients with a power consumption higher than 10 kW and below or equal to 15 kW.

- The "type b" clients are those who have contracted 2.0 DHA or 2.1 DHA tariffs. These are low voltage tariffs with time-of-use (TOU) pricing, in two daily periods: peak and valley. The 2.0 DHA tariff is for clients with a power consumption below or equal to 10 kW, and the 2.1 DHA tariff is for clients with a power consumption higher than 10 kW and below or equal to 15 kW. These TOU tariffs begun to be applied in Spain on the first of January, 2008.

- The "type c" clients are those who have contracted 3.0 A or 3.1 with low voltage measure tariffs. These are TOU tariffs for clients with a contracted power higher than 15 kW. Tariff 3.0 A is for low voltage clients (< 1000 V) and tariff 3.1 is for high voltage clients (≥ 1000 V). The defined periods for pricing are three: peak, valley and flat.

According to the information from the Spanish National Commission of Energy (CNE) from the year 2009, type *a*, type *b* and type *c* customer numbers in Spain were the following:

- Type *a* customers: 24.835.412 ($92,98\%$ from total)

- Type b customers: 1.165.001 ($4,36\%$ from total)
- Type c customers: 709.276 ($2,66\%$ from total)

Of a total of 26.709.689 customers. Adequate samples that represent in percentage the total population [26] can be computed applying the formula described in Eq. (1), where n, N, t, p and e stand for:

- n: is the sample size, that represents the population percentage.
- N: is the total population number, for a known, finite size.
- t: confidence interval parameter, obtained as a given value of confidence α. A usual value of α is $0,05$, or the 95% of confidence. For this value, the value of the t parameter equals $1,96$.
- p: is the expected proportion in the sample, i.e., $92,98\%$ in the case of type a, $4,36\%$ in the case of type b, and $2,66\%$ in the case of type c.
- e: is the expected error in the sample. A usual value is to assume an error in the sample of 3%, therefore a value of $e = 0.03$ has been used.

$$n = \frac{Nt^2p(1-p)}{(N-1)e^2 + t^2p(1-p)} \tag{1}$$

Assuming the preceding values, the resulting sample sizes for the three types of clients are:

- Type a: sample of 279 clients.
- Type b: sample of 178 clients.
- Type c: sample of 111 clients.

The database analyzed in the present work is comprised of load profiles from 759 clients, being 711 of them of type a, 44 of type b and 4 of type c. From these results it can be concluded that the sample is adequate for clients of type a, and inadequate for clients of types b and c, if no other stratification variables are taken into account, such as the presence of sufficient samples from the different climate regions of Spain. However, this information is not used for the segmentation, but to extract conclusions from its results.

Table 2: Objects data set. Description of columns or variables.

2.1.2. Database description

The main values used in this work are defined in Table 2. These data comprise the user ID, the 24 hourly energy consumptions per day, and season and day indices. Other information is also available, such as the climate area for each client, and a set of qualitative and quantitative indices added. These indices describe the characteristics of the user and the load profile, and were obtained by means of questionnaires, submitted to the population sample. These questionnaires ask the sample of clients about their home, type of building, number of residents, appliances owned, and consumption habits. The results have been computed to assign values to certain indices. These indices comprise the active management possibilities (percentage of high, medium and low penetration appliances at home, estimation of potentially manageable power, expressed in kWh per week); the familiar and residential characteristics (number of children, characteristics of the building); and the concern with environmental issues, the renewable energies and the knowledge about the current fee of electrical consumption. Although these data have not been used in the clustering process, they are available and can be obtained by the user ID, allowing further analyses regarding the users' status, geography and habits and the relation with the pattern of load profile. This work can be found in other articles, such as [27].

2.2. Description of the K-means dynamic clustering algorithm

In this paper, a modified K-means clustering is used to perform the dynamic clustering on the load profiles database. The K-means algorithm is one of the most known and used clustering algorithms, due to its efficiency and robustness. The name refers to the number k of clusters to be found, defined as an initial parameter. Following, the procedure of the classic, standard algorithm is detailed. Next, the modifications added to perform the dynamic data mining are explained.

The steps for the K-means clustering algorithm to evaluate the data are the following:

1. Select k objects and set them as the initial prototypes of the k clusters that are to be found. This can be done at random from the data set. However, this choice can delay the execution time of the algorithm and

affect its performance. Therefore other options are suggested instead, such as setting the initial prototypes based on heuristics or knowledge of the data to be clustered; or to apply specific algorithms to initialize centroids, such as the *K-means* ++ algorithm [28].

2. Compute all the Euclidean distances of the remaining objects to the k prototypes. Assign each object to the cluster with the smallest distance.
3. Compute clusters prototypes or centroids as the average or mean value from all the objects that belong to the cluster, with the objective to minimize the cost index shown in (2). This index is a summation of all the k summations of the distances from all the objects to the centroid of each cluster.
4. Proceed with the two previous steps until a termination condition is reached, like the variation in centroids falling under a predefined limit.

$$J = \sum_{i=1}^{k} \left(\sum_{j, z_j \in A_i} \| z_j - c_i \| \right) \tag{2}$$

The efficiency of the K-means algorithm highly relies on the parameter of the number of clusters (k) to be found, and its adequacy to the real number of clusters. The K-means applies the Euclidean distance as a similarity metric. This distance is a particular case of the Minkowski metric, whose definition is given in Eq. (3).

$$d(i,k) = \left(\sum_{j=1}^{d} |x_{ij} - x_{kj}|^r \right)^{\frac{1}{r}} \text{, where } r \geq 1 \tag{3}$$

The Minkowski metric is different as a function of the r parameter. The most used Minkowski metrics are three: the *Manhattan* distance ($r = 1$), the *Euclidean* distance ($r = 2$) and the *Chebyshev* distance ($r \to \inf$). The mathematical definition of the Euclidean distance can be seen in Eq. (4).

$$d(i,k) = \left(\sum_{j=1}^{d} |x_{ij} - x_{kj}|^2 \right)^{\frac{1}{2}} = \sqrt{(x_i - x_k)^T (x_i - x_k)} = \|x_i - x_k\| \tag{4}$$

Concerning the dynamic clustering, the objective of performing dynamic segmentation on a time series database is to obtain dynamic centroids, i.e., patterns that represent a number of objects whose features may vary with time, but remain similar enough to pertain to the same cluster. The objective, therefore, is to obtain a set of patterns that depict the full evolution of the data through time.

The clustering technique used must deal with objects that may have time-series discontinuities, i.e., gaps between time measures, which can vary among the different objects in the database. This is significantly true for daily load curves: measures from the meters can be unavailable for some days, due to unknown reasons, such as a malfunction or maintenance. These discontinuities do not happen with the same frequency, nor at the same days for all the clients, therefore occasional, unpredicted discontinuities appear at the data, which does not necessarily mean that a client has been removed from the database, simply that there are not available measures. Clients in this situation should be kept at their last state within a cluster, until a new measure is analyzed.

In the case of dynamic clustering, or trajectories clustering with fixed classes (type 3 in Table 1), Weber [25] describes the algorithm called Functional Fuzzy C-means or FFCM. This algorithm is presented as a generalization of the static fuzzy c-means or FCM [29]. The distance between two trajectories, f and g, is computed in three steps:

1. The fuzzy Membership Function (MF) "approximately zero" is defined ($\mu(f(x))$). This is a symmetric MF, with the maximum membership (1) at zero value, that can have any shape. Weber defines a gaussian MF, which rapidly decreases as the value of x increases.
2. Compute the *similarity* function $s(f,g)$ between two trajectories f and g as the fuzzification of the difference between the two trajectories, i.e., $s(f,g) = \mu(f(x) - g(x))$.
3. Compute the distance between the two trajectories f and g as the inverse of the similarity: $d(f,g) = (1/s(f,g)) - 1$.

In this paper, however, a dynamic K-means clustering algorithm has been developed, by modifying the static K-means algorithm to obtain the similarity distances among objects taking into account all the Euclidean distances between each pair of objects from their coincident time stamps. The process diagram of this computation is illustrated in Fig. 3. A number of analyses has been performed applying this algorithm, varying the data to be clustered. The results are described in the following section.

Figure 3: Computation of distances objects – clusters.

3. Dynamic clustering segmentation

When addressing the dynamic segmentation of the daily load profiles database, different options of granularity can be applied, according to the objectives of the segmentation. Some of these options are, for instance, the following:

1. As n sequential daily load curves.
2. As cumulated or average n weekly or monthly load curves, to avoid weekly variations due to working days.
3. As n sequential daily load curves arranged by type of day, to avoid weekly variations due to working days (e.g., all the weekend daily curves only, all the working day curves only, etc.)

In this paper the three analyses have been performed. Four different segmentation results are described, which are the following:

- Total sequence of daily load profiles through years 2008 and 2009.
- Sequence of working days daily load profiles through years 2008 and 2009.
- Sequence of non-working days daily load profiles through years 2008 and 2009.
- Sequence of cumulated monthly daily load profiles through years 2008 and 2009.

The non-working days data set in this analysis is comprised by all the Saturdays and Sundays of years 2008 and 2009, plus the holidays according to the Spanish calendar (e.g., the 1^{st} of May). The data set of working days includes all the other days.

Regarding the number of clusters to be found, the present analysis has been performed using a number of 10 clusters. This choice for the number of clusters is based on a previous work from the authors [27], where clustering

Figure 4: Cluster prototypes from dynamic clustering on sequence of daily load profiles.

Table 3: Clusters obtained from dynamic clustering on sequence of daily load profiles.

and classification techniques are applied on the energy consumption daily load profiles and the *DB index* [18] is computed for a scope of cluster numbers. However, further studies could be made in this sense, regarding the adequacy of the number of clusters chosen and the definition of indicators that value the partition performed.

3.1. Dynamic clustering on sequence of daily load profiles through years 2008 *and* 2009

The first analysis is the dynamic clustering performed on the sequential daily load profiles of the full sample of 759 clients through years 2008 and 2009, therefore the resulting clusters include 759 clients x (365 + 364) = 554.829 load profiles (the days March, the 30^{th}, 2008 and March, the 29^{th}, 2009 were lost due to the pre-treatment process of the data). The resulting centroids are depicted in Fig. 4.

Table 3 highlights some of the main values that can be obtained by observation of the resulting clusters. This information can also be easily extracted by means of mathematical functions or processes. The dynamic clustering indicates that most of the clients belong to a group of a low level of energy consumption (maximum consumption of 500 Wh) and an expectable pattern of valley and peak hours through the day. Some clients present a higher level of consumption (clusters 2 and 5), and other clients are grouped under patterns of high energy consumption during the night, at the beginning of 2008, that have experienced a shift towards the first hours of the morning in 2009. The reason for this shift might be mainly originated by the extinction of the nocturnal tariff in 2008, which would be used by clients to cumulate energy during the night, to be consumed during the day.

3.2. Dynamic clustering on sequence of daily load profiles through years 2008 *and* 2009*, working days*

The following analysis has been performed on the sequential daily load profiles, only for the working days through years 2008 and 2009, therefore

Figure 5: Cluster prototypes from dynamic clustering on sequence of daily load profiles, working days.

Table 4: Clusters obtained from dynamic clustering on sequence of daily load profiles, working days.

the resulting clusters include 759 clients x 516 working days = 391.644 load profiles. The resulting centroids are depicted in Fig. 5.

Table 4 indicates some of the main values from the resulting clusters, obtained as a first view of the cluster centroids. Again, this information could also be easily extracted by means of mathematical functions or processes. The dynamic clustering on the separate groups of load profiles, in working days and non-working days, has been performed in order to remove the known distinction in load profiles patterns between these two types of days. This way, the analysis of the differences found in patterns will not include the variability due to the sequence of working and non-working days through the year. As initial results from the observation of the centroids, similar results to the previous analysis that includes all the daily load profiles can be found. However, since no variability due to non-working days is present, the analysis yields also more clear differences in characteristic load profile patterns among the sample of clients: most of them belong to the group of low level of energy consumption and the usual load profile for residential users (cluster 7). Another group with a lower number of clients depicts the same usual pattern of load profile but a higher level of energy consumption, therefore being an appropriate objective for DSM actions (cluster 4). There are also groups with high energy consumption at nighttime (clusters 3, 9 and 10), and also clients with an unusual pattern of load profile and level of energy consumption (clusters 2 and 5). The seasonality effect is observed among all the patterns.

3.3. Dynamic clustering on sequence of daily load profiles through years 2008 *and* 2009*, non-working days*

The following analysis has been performed on the sequential daily load profiles, only for the non-working days through years 2008 and 2009, therefore the resulting clusters include 759 clients x 213 non-working days = 161.667 load profiles. The resulting centroids are depicted in Fig. 6.

Figure 6: Cluster prototypes from dynamic clustering on sequence of daily load profiles, non-working days.

Table 5: Clusters obtained from dynamic clustering on sequence of daily load profiles, non-working days.

Table 5 indicates some of the main values from the resulting clusters, obtained as a first view of the cluster centroids. Load profile patterns from non-working days differ from working days in levels and hours of energy consumption. It can be seen that the main group of clients (cluster 5) displays a pattern of low level of energy consumption and a profile of peak and valley hours slightly different from that of working days. The second group with the highest number of clients (cluster 10) has a higher consumption of energy through the day, and the peak hours after lunch approximate in energy consumption those of the energy consumed after dinner. The group of users with a high level of energy consumption at night is also observed (cluster 8), but also some patterns are obtained, of clients that consume energy at an approximately constant rate through the day (clusters 3, 4, 6 and 9).

3.3.1. Dynamic clustering on sequence of cumulated monthly daily load profiles through years 2008 *and* 2009

The last of the analyses performed has been done on the aggregated or cumulated hourly energy used at the end of each month, from all the customers, for years 2008 and 2009, therefore the resulting clusters include 759 clients $x24$ months = 18.216 load profiles. The resulting centroids are depicted in Fig. 7.

Table 6 indicates some of the main values from the resulting clusters, obtained as a first view of the cluster centroids.

From the results it can be seen that there are 2 clusters with the majority of clients that display the typical load profile of peaks and valley hours, with slight differences among them, and different levels of average energy consumption. These would correspond to clusters 7 and 8. In clusters 4 and 6, other type of electricity consumption patterns can be observed, with a more clear trend on an increase in consumption through the day. Clusters 2 and 9 gather a reduced group of users (10) with a high rate of energy consumption during the night. Besides, a shifting trend can be observed in

Figure 7: Cluster prototypes from dynamic clustering on sequence of cumulated monthly daily load profiles.

Table 6: Clusters obtained from dynamic clustering on sequence of cumulated monthly daily load profiles.

energy consumption in the group of users from cluster 9, from a peak in the first hours of the day in the first months of 2008, to a more distributed consumption pattern in the last months of 2009. Finally, clusters 1, 5 and 10 represent 45 users with different patterns of energy consumption. A further analysis by an expert could be made to determine the nature for the obtained patterns.

4. Results and Discussion

Seasonality effects can be clearly observed, with generalized higher consumptions in winter, lower energy consumptions in summer, and the lowest values from autumn and spring. It is also interesting to analyze whether changes in the Spanish legislation of the energy market have affected the way residential clients make use of the energy; these changes would be reflected along the dynamic patterns of the resulting centroids. However, no significant change is observed. From July, the first, 2009, low voltage users with contracted power equal or below 10 kW were given the choice to either be charged under a specific, fixed tariff (TUR), or to contract the energy with an authorized retailer in a liberalized energy market. Low voltage users with contracted power higher than 10 kW and every high voltage customer had to contract their energy tariffs in the liberalized energy market. The change in legislation has mainly influenced the behavior of the clients from the disappeared nocturnal tariff, which has been switched to mostly type b users. These are the clusters of clients with high energy consumption at night hours. It can be observed from the results that in the patterns from some clusters the loads have shifted from night hours to morning or midday hours.

Finally, it can be observed that groups of clients with higher energy consumption than the average are grouped at the same clusters. These users should be the main objective for DSM actions, since it is more likely that manageable equipment and appliances can be found at these residences.

The dynamic clustering allows capturing the trend of groups of users at a glance. As can be seen in the results, the clients are clustered by level of energy consumption and by the form of their load profiles, thus allowing a classification of these users according to the way they consume energy (when and how much).

5. Conclusions

The same dynamic clustering analysis, based on the K-means algorithm, has been performed on 4 data sets of the same time series, a database of electricity consumption from residential consumers in Spain. The following conclusions can be drawn from these analyses:

- The results obtained in the four cases allow a fast identification of the main types of energy consumption patterns in the group of Spanish residential users of electric energy. This distinction can be more easily observed in the first three analyses, rather than in the cumulated monthly values of the energy consumed by each hour. Three main types of energy consumption users have been identified, which are explained next.

- The first type of client gathers the majority of the users in the sample (around 700 clients), and also represent the common pattern of energy consumption in domestic or residential users in Spain. It is represented by $2 - 4$ clusters in each of the four analysis, which represent a daily profile with three ascending peaks of energy consumption: one in the morning (around 8), another one at lunch (around 15 h.) and the highest one at night, around 22 h. The dynamic clustering groups these clients according to the shape of these peaks and the level of energy consumption, typically representing clients of low and medium energy consumption ($500 - 1.500$ Wh maximum). These results can be observed in the four analyses, but better in the first three. In the first analysis, for instance, this group is represented by clusters 1, 3 and 10.

- The second type of clients represents a minority of users with a high level of energy consumption through the day. These clusters can also be observed in the four analyses, however it is best observed in the first three. The clusters obtained have two main different shapes: one with the typical shape of energy consumption, described above, but with higher energy levels (from 2.500 to 7.000 Wh), for instance clusters 2, 5, 7 and 8 in the first analysis; and another group of users that present a (more or less) flat shape of energy consumption through the day (including the night hours), for instance cluster 9 in the first analysis.

- The third type comprehends clients with a higher consumption of energy at night. Examples of these clusters in the first analysis are clusters 5 and 6.

- Finally, in the first three analysis, there can be observed that the clustering process has identified 1 client with an anomalous pattern of energy consumption for residential use, with values of energy consumption that reach up to 10.000 kWh through daylight. In the first analysis, for instance, this client can be found in cluster 8. The dynamic clustering allows, therefore, a fast identification of anomalous or unexpected patterns of energy consumption.

The dynamic clustering analysis would be an efficient tool for clients' classification and trend behavior. The objective of DSM is one of the direct applications of this technique, however not the only one. Fraud detection could also be another possibility: a quick clustering of consumers in patterns will detect and highlight specific groups of clients whose level and profile of energy consumption may not suit the terms of their contracted power and energy tariffs.

The results of the analysis could be sufficiently representative of the type *a* residential customers in Spain, if no other strata differentiation are taken into account. If other stratifications are to be taken into account, however, such as for instance, the different climate regions in Spain, the idoneity of the clients analysed to the stratified sample must be appropriately addressed.

From this analysis, an expert or operator should identify and classify objective clients. These decisions could be supported by decision support systems and an automated analysis of the resulting clusters, performing evaluation of trends, detection of anomalous behaviors, and automatically suggesting groups of clients for specific actions, such as commercial offers, or DSM orders for energy reduction. Prediction or load forecasting may also be combined with the dynamic clustering, to provide helpful estimation of patterns behaviors in medium term.

6. Acknowledgements

The works developed in this document have been possible thanks to Iberdrola Distribución Eléctrica S.A.U. and the development of the GAD project. The GAD or "Active Demand Management" (in Spanish) project was a project supported by the Spanish Government, and participated by 14 different companies and 14 research centers. It was sponsored by the CDTI (Technological Development Centre of the Ministry of Science and Innovation of Spain), and financed by the INGENIO 2010 program.

The promotion of the project came from the National Strategic Consortium of the Electrical Active Demand Management; Iberdrola Distribución Eléctrica, S.A.U. was the leader of the project, and the rest of former companies were: Red Eléctrica de España, Unión Fenosa Distribución, Unión Fenosa Metra, Iberdrola S.A., Orbis Tecnología Eléctrica, ZIV Media, DIMAT, Siemens, Fagor Electrodomésticos, BSH Electrodomésticos España, Ericsson España, GTD Sistemas de Información, Grupo Foresis - Acceda Mundo Digital and Airzone. The objective of the GAD project was to investigate and develop solutions to optimize the electrical consumption in low and medium voltage users, by the research and development of new tools for the DSM. The data analyzed has been provided by the GAD project, as a database with load profile registers from smart meters, gathered from the sample of Spanish clients selected.

References

[1] Benefits of demand response in electricity markets and recommendations for achieving them, Tech. rep., U.S. Department of Energy (2005).

[2] D. Crossley, Worldwide survey of network-driven demand-side management projects, Tech. rep., International Energy Agency - Demand Side Management Programme (2008).

[3] I. Benítez, J. L. Díez, P. Albertos, Applying dynamic mining on multiagent systems, in: Proceedings of the 17th World Congress. The International Federation of Automatic Control (IFAC). Seoul, Korea, July 6-11, 2008.

[4] F. Crespo, R. Weber, A methodology for dynamic data mining based on fuzzy clustering, Fuzzy Sets and Systems 150 (2005) 267–284.

[5] S. Mitra, S. K. Pal, P. Mitra, Data mining in soft computing framework: A survey, IEEE Transactions on Neural Networks 13 (1) (2002) 3–14.

[6] J. Han, M. Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann, 2006.

[7] D. Hand, H. Mannila, P. Smyth, Principles of Data Mining, The MIT Press, 2001.

[8] J. Jackson, Data mining: A conceptual overview, Communications of the Association for Information Systems 8 (2002) 267–296.

[9] C.-S. Chen, J. C. Hwang, C. Huang, Application of load survey systems to proper tariff design, Power Systems, IEEE Transactions on 12 (4) (1997) 1746–1751.

[10] C. Chen, M. Kang, J. Hwang, C. Huang, Synthesis of power system load profiles by class load study, International Journal of Electrical Power & Energy Systems 22 (5) (2000) 325 – 330.

[11] G. Chicco, R. Napoli, F. Piglione, P. Postolache, M. Scutariu, C. Toader, Emergent electricity customer classification, Generation, Transmission and Distribution, IEE Proceedings- 152 (2) (2005) 164–172.

[12] G. Chicco, R. Napoli, F. Piglione, Application of clustering algorithms and self organising maps to classify electricity customers, in: Power Tech Conference Proceedings, 2003 IEEE Bologna, Vol. 1, 2003, pp. 7 pp. Vol.1–.

[13] S. Verdu, M. Garcia, F. Franco, N. Encinas, A. Marin, A. Molina, E. Lazaro, Characterization and identification of electrical customers through the use of self-organizing maps and daily load parameters, in: Power Systems Conference and Exposition, 2004. IEEE PES, 2004, pp. 899–906 vol.2.

[14] S. Verdu, M. Garcia, C. Senabre, A. Marin, F. Franco, Classification, filtering, and identification of electrical customer load patterns through the use of self-organizing maps, Power Systems, IEEE Transactions on 21 (4) (2006) 1672–1682.

[15] T. Kohonen, Self-Organizing Maps, Springer, 2001.

[16] G. Tsekouras, N. Hatziargyriou, E. Dialynas, Two-stage pattern recognition of load curves for classification of electricity customers, Power Systems, IEEE Transactions on 22 (3) (2007) 1120–1128.

[17] G. Tsekouras, P. Kotoulas, C. Tsirekis, E. Dialynas, N. Hatziargyriou, A pattern recognition methodology for evaluation of load profiles and typical days of large electricity customers, Electric Power Systems Research 78 (9) (2008) 1494 – 1510.

[18] D. L. Davies, D. W. Bouldin, Cluster separation measure, IEEE Transactions on Pattern Analysis and Machine Intelligence 1 (1979) 95–104.

[19] V. Figueiredo, F. Rodrigues, Z. Vale, J. Gouveia, An electric energy consumer characterization framework based on data mining techniques, Power Systems, IEEE Transactions on 20 (2) (2005) 596–602.

[20] R. Chang, C. Lu, Load profile assignment of low voltage customers for power retail market applications, Generation, Transmission and Distribution, IEE Proceedings- 150 (3) (2003) 263–267.

[21] J. B. MacQueen, Some methods for classification and analysis of multivariate observations, in: Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, University of California, Berkeley, CA, 1967, pp. 281–297.

[22] J. L. Díez, Técnicas de agrupamiento para identificación y control por modelos locales, Ph.D. thesis, Universidad Politécnica de Valencia, in Spanish (Julio 2003).

[23] S. Haykin, Neural Networks, a Comprehensive Foundation, MacMillan, 1994.

[24] D. Gerbec, S. Gasperic, F. Gubina, Determination and allocation of typical load profiles to the eligible consumers, in: Power Tech Conference Proceedings, 2003 IEEE Bologna, Vol. 1, 2003, pp. 5 pp. Vol.1–.

[25] J. V. de Oliveira, W. Pedrycz (Eds.), Advances in Fuzzy Clustering and its Applications, John Wiley & Sons, Ltd., 2007.

[26] S. L. Lohr, Sampling: Design and Analysis, 2000.

[27] I. Sanchez, I. Espinos, L. Moreno Sarrion, A. Lopez, I. Burgos, Clients segmentation according to their domestic energy consumption by the use of self-organizing maps, in: Energy Market, 2009. EEM 2009. 6th International Conference on the European, 2009, pp. 1–6.

[28] D. Arthur, S. Vassilvitskii, k-means++: the advantages of careful seeding, in: Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms, SODA '07, Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2007, pp. 1027–1035.

[29] J. C. Bezdek, Fuzzy mathematics in pattern classification, Ph.D. thesis, Faculty of the Gradual School of Cornell University, Ithaca, NY (1973).

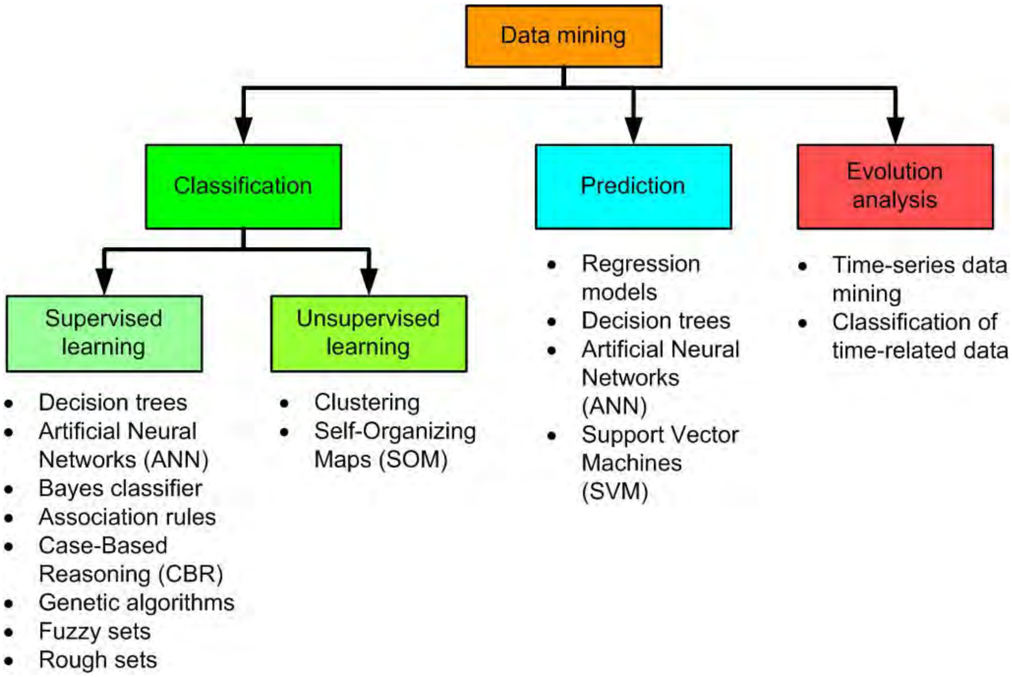

Figure 1

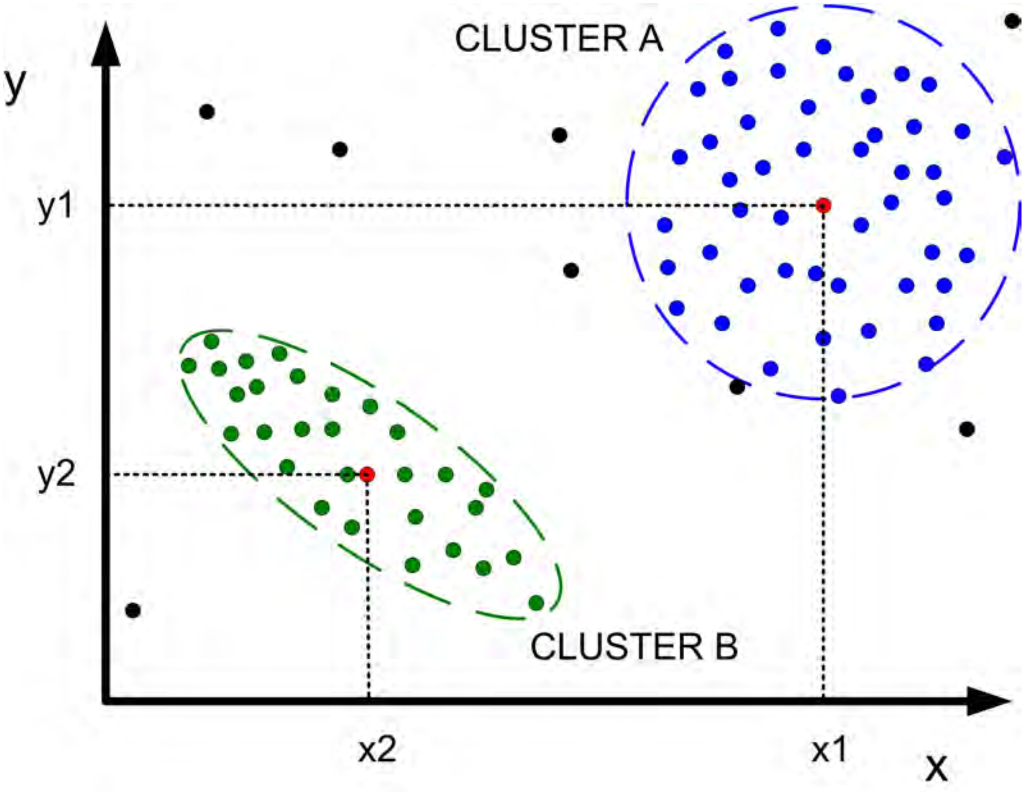

Figure 2

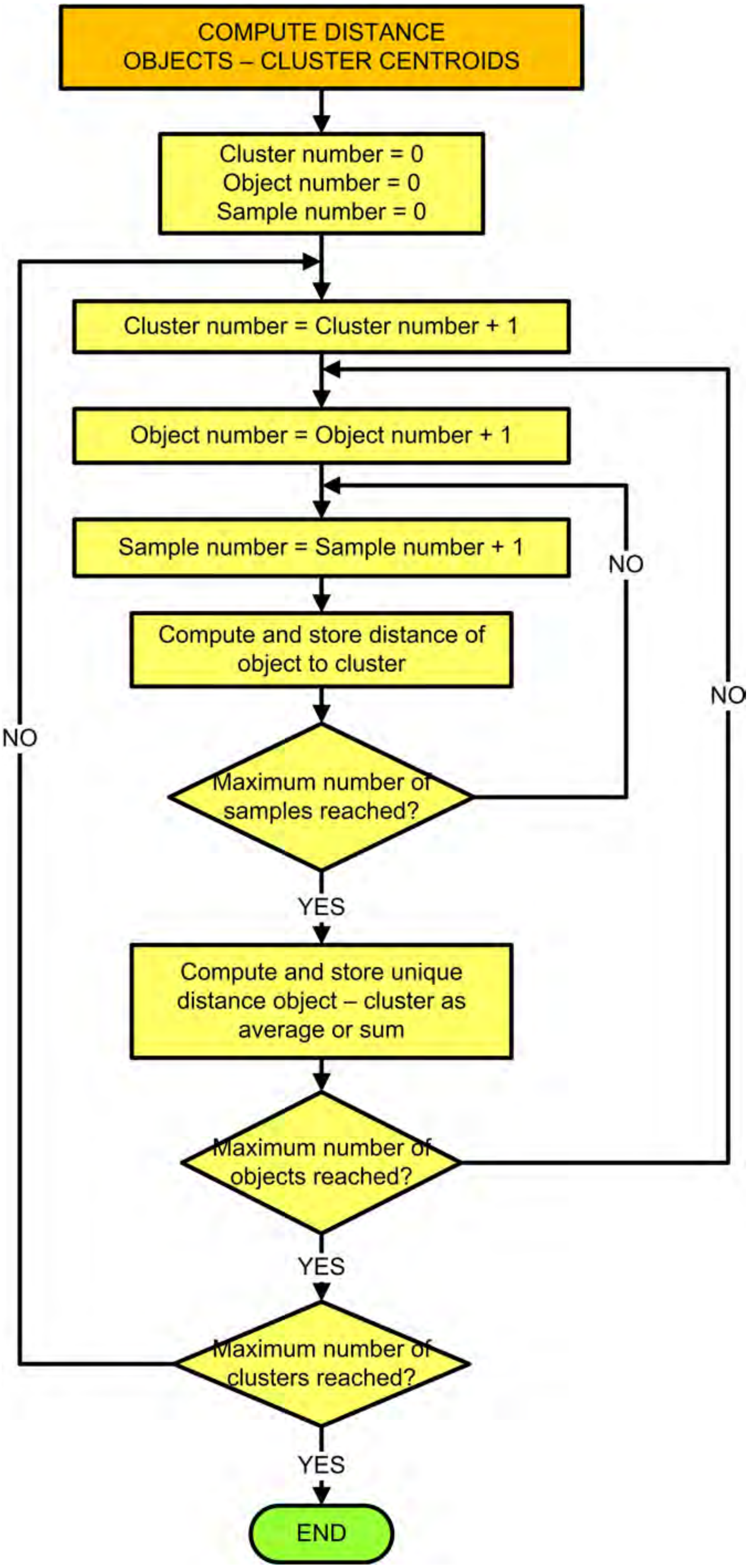
COMPUTE DISTANCE
OBJECTS – CLUSTER CENTROIDS
Cluster number = 0
Object number = 0
Sample number = 0
Cluster number = Cluster number + 1
Object number = Object number + 1
Sample number = Sample number + 1
Compute and store distance of object to cluster
Maximum number of samples reached?
NO
YES
Compute and store unique distance object – cluster as average or sum
Maximum number of objects reached?
NO
YES
Maximum number of clusters reached?
NO
YES
END

Figure 3

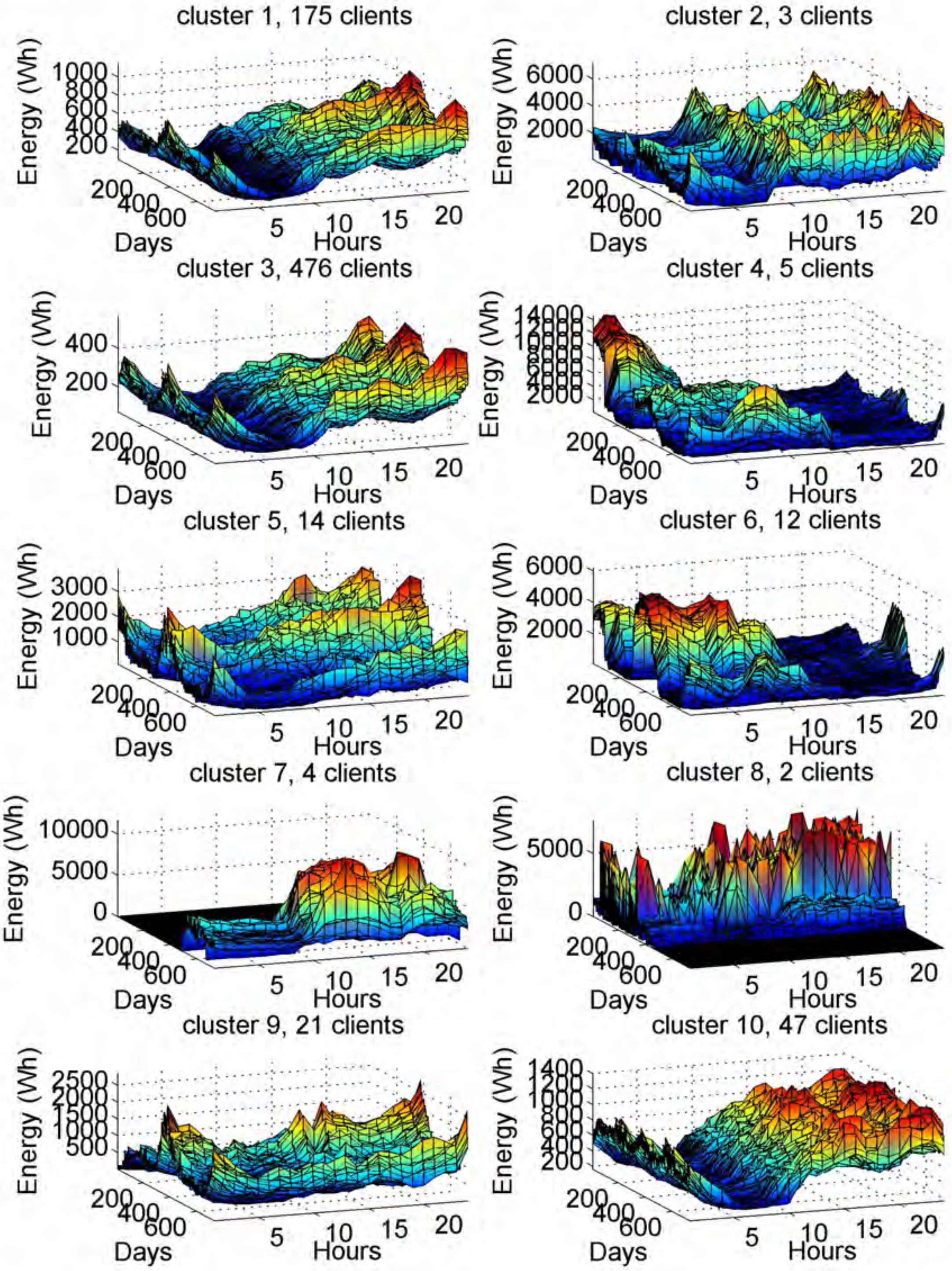
cluster 1, 175 clients
Energy (Wh)
1000
800
600
400
200
200
400
600
Days
5
10
15
20
Hours
cluster 2, 3 clients
Energy (Wh)
6000
4000
2000
200
400
600
Days
5
10
15
20
Hours
cluster 3, 476 clients
Energy (Wh)
400
200
200
400
600
Days
5
10
15
20
Hours
cluster 4, 5 clients
Energy (Wh)
14000
12000
10000
8000
6000
4000
2000
200
400
600
Days
5
10
15
20
Hours
cluster 5, 14 clients
Energy (Wh)
3000
2000
1000
200
400
600
Days
5
10
15
20
Hours
cluster 6, 12 clients
Energy (Wh)
6000
4000
2000
200
400
600
Days
5
10
15
20
Hours
cluster 7, 4 clients
Energy (Wh)
10000
5000
0
200
400
600
Days
5
10
15
20
Hours
cluster 8, 2 clients
Energy (Wh)
5000
0
200
400
600
Days
5
10
15
20
Hours
cluster 9, 21 clients
Energy (Wh)
2500
2000
1500
1000
500
200
400
600
Days
5
10
15
20
Hours
cluster 10, 47 clients
Energy (Wh)
1400
1200
1000
800
600
400
200
200
400
600
Days
5
10
15
20
Hours

Figure 4

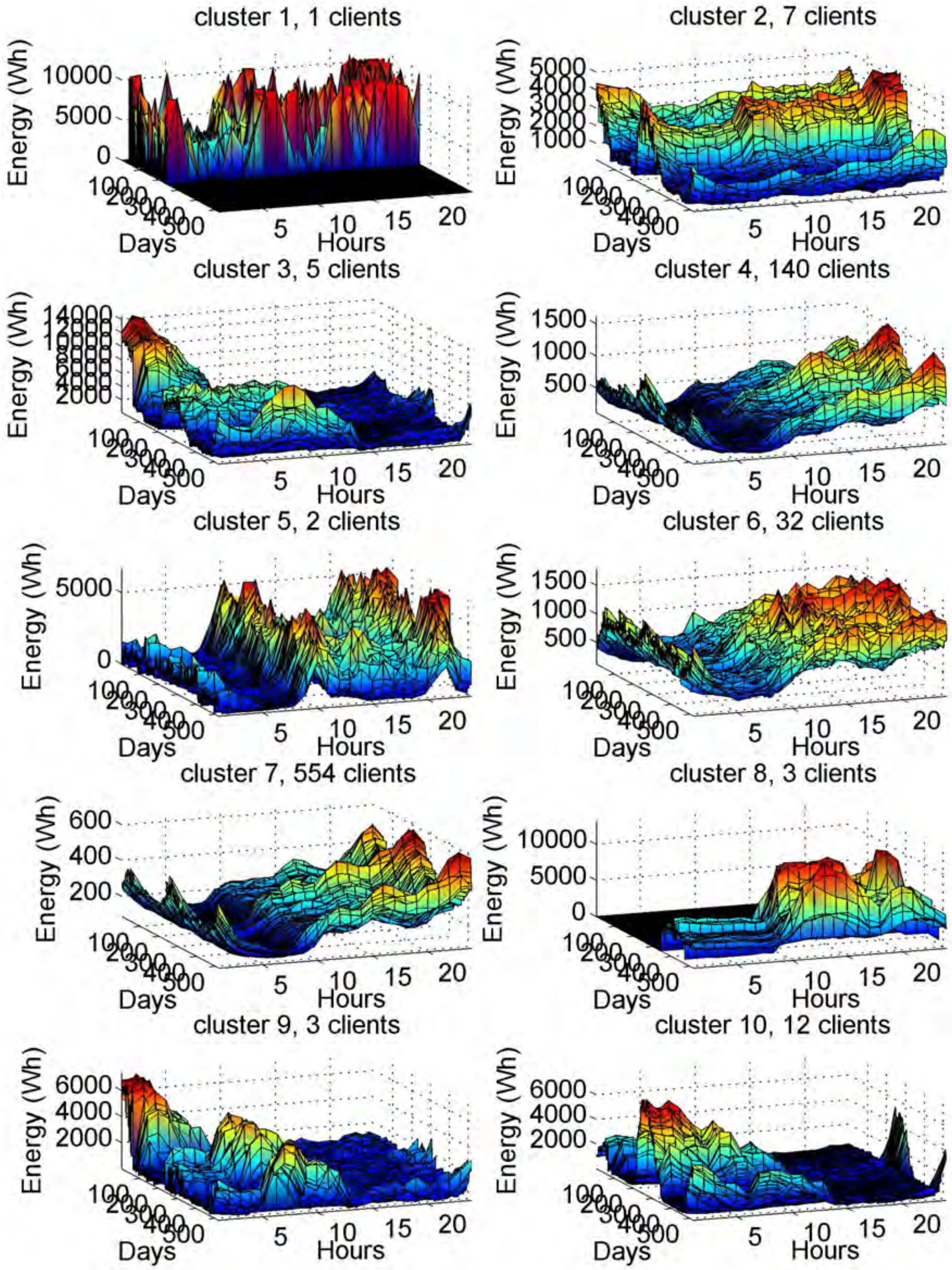
cluster 1, 1 clients
Energy (Wh)
10000
5000
0
100
200
300
400
500
Days
5
10
15
20
Hours
cluster 2, 7 clients
Energy (Wh)
5000
4000
3000
2000
1000
100
200
300
400
500
Days
5
10
15
20
Hours
cluster 3, 5 clients
Energy (Wh)
14000
12000
10000
8000
6000
4000
2000
100
200
300
400
500
Days
5
10
15
20
Hours
cluster 4, 140 clients
Energy (Wh)
1500
1000
500
100
200
300
400
500
Days
5
10
15
20
Hours
cluster 5, 2 clients
Energy (Wh)
5000
0
100
200
300
400
500
Days
5
10
15
20
Hours
cluster 6, 32 clients
Energy (Wh)
1500
1000
500
100
200
300
400
500
Days
5
10
15
20
Hours
cluster 7, 554 clients
Energy (Wh)
600
400
200
100
200
300
400
500
Days
5
10
15
20
Hours
cluster 8, 3 clients
Energy (Wh)
10000
5000
0
100
200
300
400
500
Days
5
10
15
20
Hours
cluster 9, 3 clients
Energy (Wh)
6000
4000
2000
100
200
300
400
500
Days
5
10
15
20
Hours
cluster 10, 12 clients
Energy (Wh)
6000
4000
2000
100
200
300
400
500
Days
5
10
15
20
Hours

Figure 5

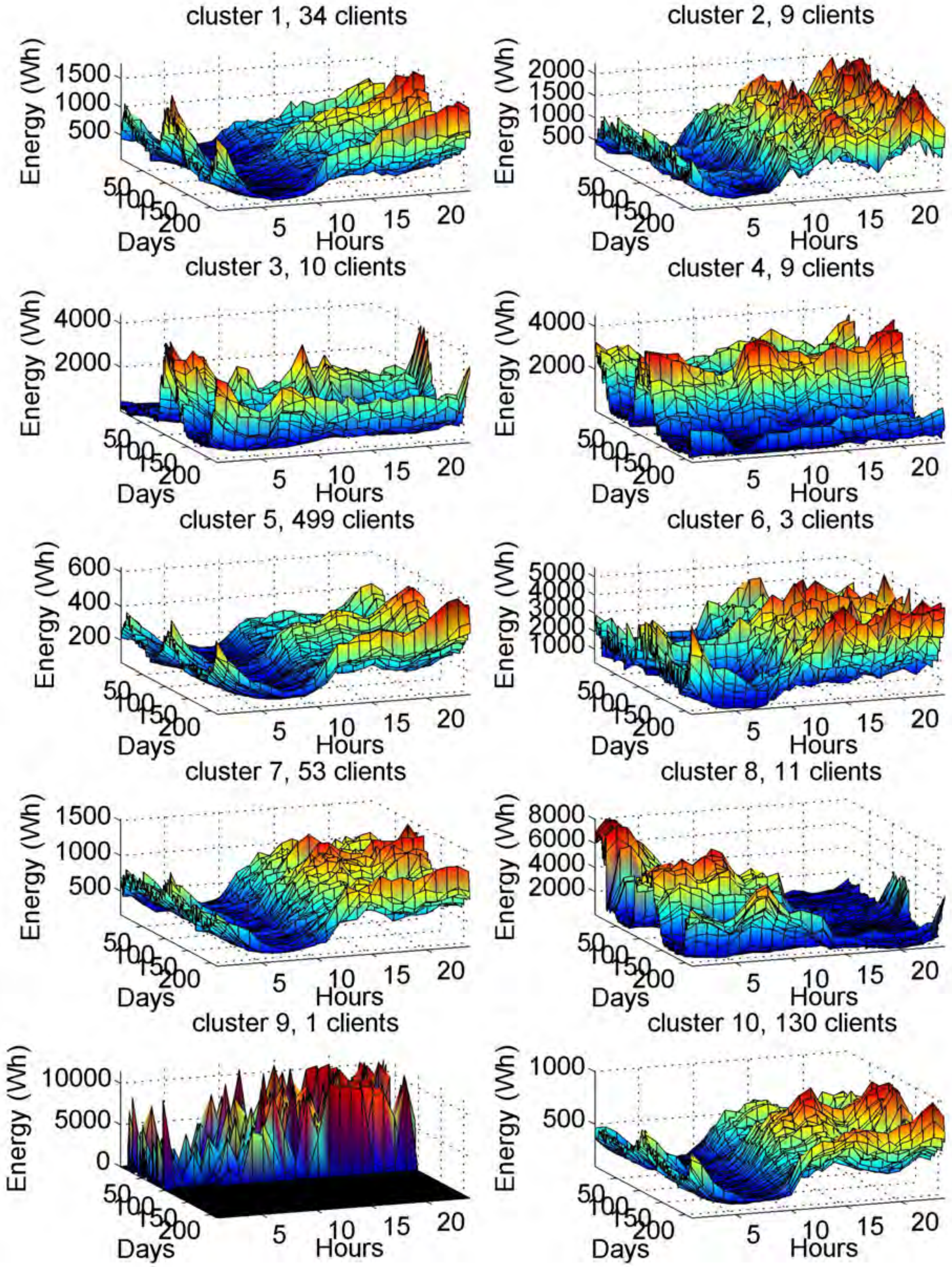
cluster 1, 34 clients
Energy (Wh)
1500
1000
500
50
100
150
200
Days
5
10
15
20
Hours
cluster 2, 9 clients
Energy (Wh)
2000
1500
1000
500
cluster 3, 10 clients
Energy (Wh)
4000
2000
cluster 4, 9 clients
Energy (Wh)
4000
2000
cluster 5, 499 clients
Energy (Wh)
600
400
200
cluster 6, 3 clients
Energy (Wh)
5000
4000
3000
2000
1000
cluster 7, 53 clients
Energy (Wh)
1500
1000
500
cluster 8, 11 clients
Energy (Wh)
8000
6000
4000
2000
cluster 9, 1 clients
Energy (Wh)
10000
5000
0
cluster 10, 130 clients
Energy (Wh)
1000
500

Figure 6

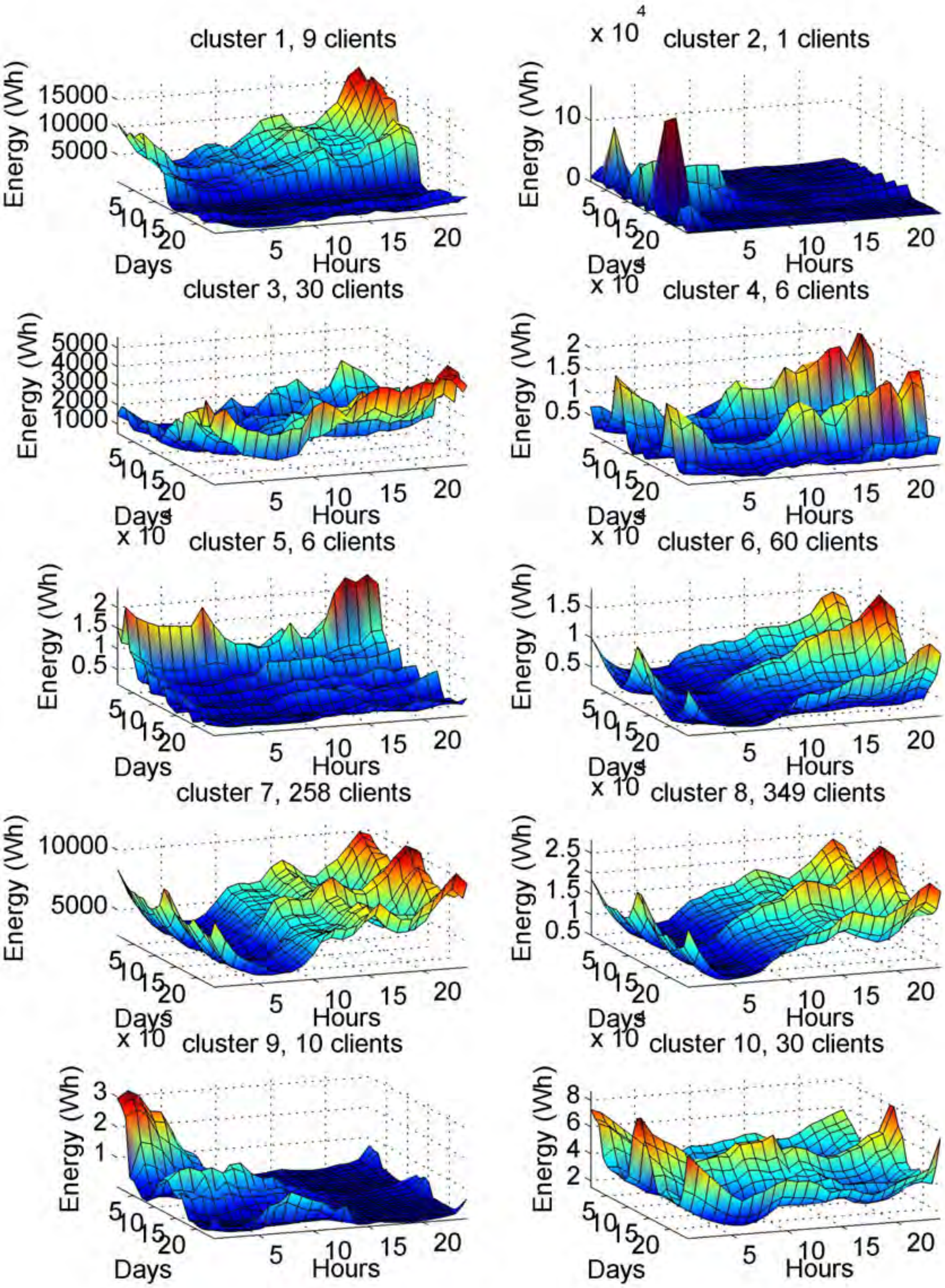
cluster 1, 9 clients
Energy (Wh)
15000
10000
5000
5 10 15 20
Days
5 10 15 20
Hours
x 10^4
cluster 2, 1 clients
Energy (Wh)
10
0
5 10 15 20
Days
5 10 15 20
Hours
cluster 3, 30 clients
Energy (Wh)
5000
4000
3000
2000
1000
5 10 15 20
Days
5 10 15 20
Hours
x 10^4
cluster 4, 6 clients
Energy (Wh)
2
1.5
1
0.5
5 10 15 20
Days
5 10 15 20
Hours
x 10^4
cluster 5, 6 clients
Energy (Wh)
2
1.5
1
0.5
5 10 15 20
Days
5 10 15 20
Hours
x 10^4
cluster 6, 60 clients
Energy (Wh)
1.5
1
0.5
5 10 15 20
Days
5 10 15 20
Hours
cluster 7, 258 clients
Energy (Wh)
10000
5000
5 10 15 20
Days
5 10 15 20
Hours
x 10^4
cluster 8, 349 clients
Energy (Wh)
2.5
2
1.5
1
0.5
5 10 15 20
Days
5 10 15 20
Hours
x 10^5
cluster 9, 10 clients
Energy (Wh)
3
2
1
5 10 15 20
Days
5 10 15 20
Hours
x 10^4
cluster 10, 30 clients
Energy (Wh)
8
6
4
2
5 10 15 20
Days
5 10 15 20
Hours

Figure 7

Table 1

Data objects	**Classes**	**Type of clustering**
Static	Static	1. Static clustering
Static	Dynamic	2. Dynamic clustering. Prototypes and classes are updated at each cycle
Dynamic	Static	3. Dynamic clustering. Prototypes represented by feature trajectories. Fixed classes
Dynamic	Dynamic	4. Dynamic clustering. Prototypes represented by feature trajectories, classes updated iteratively

Table 2

Column number	**Description**	**Format**
1	User or client unique ID	String
2	Date	Date
3 to 26	Hourly energy consumption (Wh)	Numeric
27	Season (Summer or Winter only)	Numeric
28	Weekday	String
29	Weekday	Numeric (1 = Sunday)
30	Working day	Numeric (1 = working day, 2 = non-working day)

Table 3

Cluster no.	**No. of clients**	**Maximum energy value (Wh)**	**time of maximum energy consumption (hours)**
1	175	$\simeq 1000$	$22 - 23$
2	3	$\simeq 5000$	$20 - 21$
3	476	$\simeq 500$	$22 - 23$
4	5	$\simeq 14000$	$00 - 01$
5	14	$\simeq 3500$	$22 - 23$
6	12	$\simeq 5000$	$00 - 01$
7	4	$\simeq 8000$	$21 - 22$
8	2	$\simeq 7000$	$14 - 15$
9	21	$\simeq 2500$	$23 - 00$
10	47	$\simeq 1300$	$22 - 23$

Table 4

Cluster no.	**No. of clients**	**Maximum energy value (Wh)**	**time of maximum energy consumption (hours)**
1	1	$\simeq 10000$	$01, 11, 22 - 23$
2	7	$\simeq 5000$	$22 - 23$
3	5	$\simeq 14000$	$00 - 01$
4	140	$\simeq 1500$	$22 - 23$
5	2	$\simeq 5000$	$10, 22 - 23$
6	32	$\simeq 1600$	$22 - 23$
7	554	$\simeq 600$	$22 - 23$
8	3	$\simeq 11000$	$11 - 16$
9	3	$\simeq 7000$	$00 - 01$
10	12	$\simeq 6000$	$00 - 01$

Table 5

Cluster no.	**No. of clients**	**Maximum energy value (Wh)**	**time of maximum energy consumption (hours)**
1	34	$\simeq 1500$	$22 - 23$
2	9	$\simeq 2000$	$21 - 22$
3	10	$\simeq 4000$	$00 - 01$
4	9	$\simeq 4000$	$22 - 23$
5	499	$\simeq 600$	$22 - 23$
6	3	$\simeq 5000$	$22 - 23$
7	53	$\simeq 1400$	$22 - 23$
8	11	$\simeq 8000$	$00 - 01$
9	1	$\simeq 10000$	$11 - 16$
10	130	$\simeq 1000$	$22 - 23$

Table 6

Cluster no.	**No. of clients**	**Maximum energy value (MWh) x 1 month**	**time of maximum energy consumption (hours)**
1	9	$\simeq 16$	$21 - 22$
2	1	$\simeq 150$	$01 - 02$
3	30	$\simeq 5$	$21 - 22$
4	6	$\simeq 20$	$22 - 23$
5	6	$\simeq 20$	$22 - 23$
6	60	$\simeq 18$	$21 - 22$
7	258	$\simeq 18$	$22 - 23$
8	349	$\simeq 25$	$22 - 23$
9	10	$\simeq 300$	$00 - 01$
10	30	$\simeq 300$	$23 - 00$