Classification, Filtering, and Identification of Electrical Customer Load Patterns Through the Use of Self-Organizing Maps

Sergio Valero Verdú, Mario Ortiz García, Carolina Senabre, Antonio Gabaldón Marín, *Member, IEEE*, and Francisco J. García Franco

***Abstract*—Different methodologies are available for clustering purposes. The objective of this paper is to review the capacity of some of them and specifically to test the ability of self-organizing maps (SOMs) to filter, classify, and extract patterns from distributor, commercializer, or customer electrical demand databases. These market participants can achieve an interesting benefit through the knowledge of these patterns, for example, to evaluate the potential for distributed generation, energy efficiency, and demand-side response policies (market analysis). For simplicity, customer classification techniques usually used the historic load curves of each user. The first step in the methodology presented in this paper is anomalous data filtering: holidays, maintenance, and wrong measurements must be removed from the database. Subsequently, two different treatments (frequency and time domain) of demand data were tested to feed SOM maps and evaluate the advantages of each approach. Finally, the ability of SOM to classify new customers in different clusters is also examined. Both steps have been performed through a well-known technique: SOM maps. The results clearly show the suitability of this approach to improve data management and to easily find coherent clusters between electrical users, accounting for relevant information about weekend demand patterns.**

***Index Terms*—Data mining, demand management, electrical customer segmentation, load patterns, self-organizing maps (SOMs).**

I. Introduction

THE liberalization process of the electrical market has not been as successful as was planned, due to a lot of problems that have appeared since 2000 until now: for example, the California energy crisis in 2000 or blackouts in Europe, the United States, and Canada in 2003. Due to these experiences, regulators and system operators believe more and more that additional electricity resources (distributed energy resources) should be procured using an integrated process that would take into account not only supply but also demand policies: for example, efficiency gains in demand (in a long-term horizon), demand management, or price responsiveness (in short-term horizon). The effective contribution to these programs and the necessity of offering energy choices to consumers need: a detailed knowledge of customer segments, the characterization of these segments (demand behavior), end-uses "dissection" for each customer, load modeling (demand and response models), and further demand aggregation to achieve demand packages for demand-side biddings and offers in energy markets (see Fig. 1).

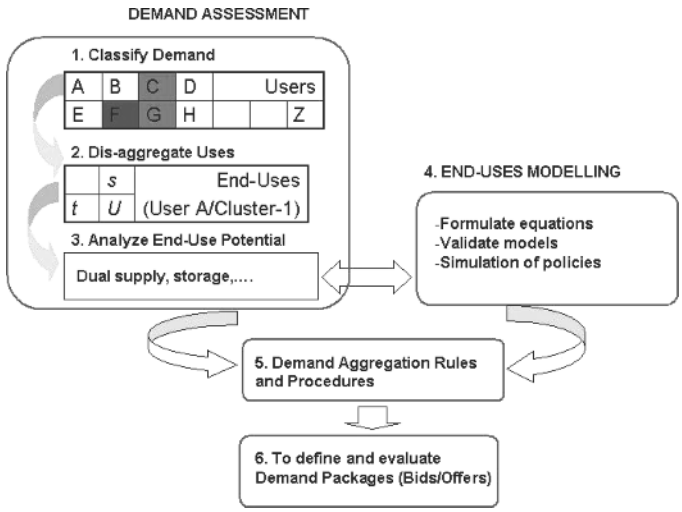

Fig. 1. Methodology to analyze, evaluate, and enhance the possibility of demand participation (DSB/DR) in electricity markets.

Besides, this deregulation and liberalization in power systems caused the necessity of new (customer and system) measurements, monitoring, and control activities. This fact has increased the amount of data stored by supply-side actors. So, this enormous quantity of available data presents a problem for utilities but also a non-negligible opportunity for distribution research. This high-dimensional data set cannot be easily modeled, and advanced tools for synthesizing structures from such information are needed.

Previous results on modeling, aggregation, and construction of energy packages were presented by the authors of [1] and [2]. The rest of this paper presents a methodology for customer segmentation and classification through the improvement and use of the data mining or knowledge discovery in databases techniques [3], [4].

Manuscript received October 4, 2005; revised May 5, 2006. This work was supported by European Union Sixth Framework Program under Project EU-DEEP SES6-CT-2003-503516. Paper no. TPWRS-00633-2005.

S. V. Verdú and M. O. García are with the Department of Electrical Engineering, Universidad Miguel Hernández, Elche, Spain (e-mail: svalero@umh.es).

C. Senabre is with the Department of Mechanics, Universidad Miguel Hernández, Elche, Spain.

A. Gabaldón Marin and F. J. G. Franco are with the Department of Electrical Engineering, Universidad Politécnica de Cartagena, Cartagena, Spain (e-mail: antonio.gabaldon@upct.es).

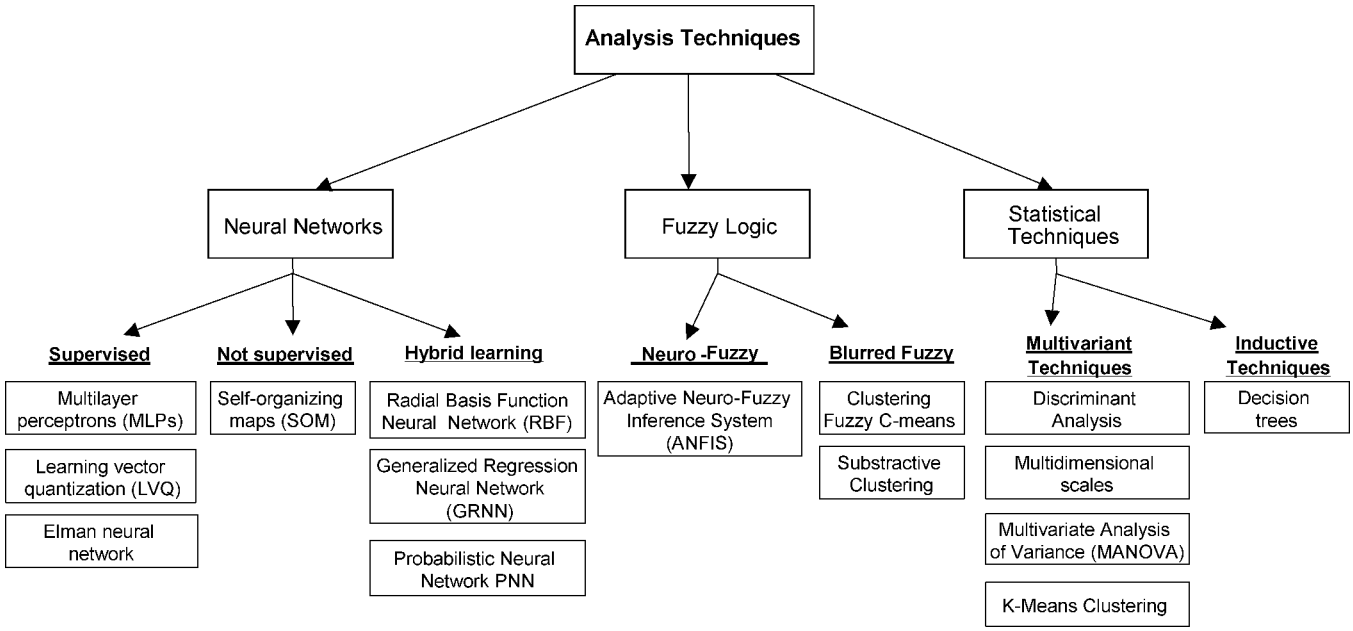

Fig. 2. Analysis techniques.

II. Review of Customer Classification Methodologies

In some data mining tutorials [3], classification methodologies are grouped in different categories according to the main task they are usually focused on: artificial intelligence techniques (neural networks and fuzzy logic), statistical techniques (linear regression and discriminant analysis), and visualization techniques (histograms, dendograms, and scatter plots). Fig. 2 shows a compendium of the techniques mentioned above and tested for this paper.

A. Techniques Review

The following paragraphs describe the characteristics of the most interesting methodologies presented in Fig. 2.

1) Artificial Neural Networks Techniques: Artificial neural networks (ANNs) try to reproduce the way the human brain acts: a highly complex, nonlinear, and parallel information processor able to perform certain computations many times faster than the most powerful digital computer available today.

Actually, ANNs find applications in such diverse fields as modeling, time-series analysis pattern recognition, and others by virtue of their ability to learn from input data with or without a teacher.

First important results in the ANN field were obtained with the simple perceptron (1958) [5] and the adaptive linear element (ADALINE) (1960), two supervised learning neural networks able to classify linearly separable sets of vectors.

Simple perceptron evolved into multilayer perceptrons (MLPs), feedforward neural networks with more than one perceptron used to solve more difficult problems.

Later, in the 1980s, Kohonen introduced the learning vector quantization (LVQ) [6] based on competitive layers in which neurons compete with each other for the right to respond to a given input vector: individual neurons learn to become feature detector cells.

Finally, Elman networks [7] are able to learn, recognize, and generate temporal patterns, as well as spatial patterns, by means of the recurrent connection feature of the network.

If the target outputs are not available, unsupervised networks must be used. In this case, the weights and biases of the network are only modified in response to inputs (so target outputs are not needed), and the algorithms classify the input patterns in a finite number of classes.

Self-organizing maps (SOMs) [6] are unsupervised networks able to learn both the distribution (as competitive layers do) and the topology of the input vectors on which they are trained. Consequently, excellent clustering results are obtained. In addition, an easy evaluation of the result is possible through the graphical representation on maps whose different labels (customers or vectors identifiers) can be grouped by visual inspection. Applying some index functions, it is possible to obtain an optimum clustering, but some "supervision" is necessary to filter the results of the maps (i.e., the operator selects the maximum number of clusters). More detailed information is presented in Section III.

The main features of the supervised and unsupervised techniques discussed above can be consulted in Table I.

Some methodologies in Fig. 2 appear as "hybrid learning" techniques. A hybrid method for learning encompasses two phases: the first is a not supervised one for the determination of clusters center, and the second is a supervised phase, for the weights and thresholds determination [8].

Three different techniques are presented: radial basis networks [9], generalized regression neural networks (GRNN) [10], and probabilistic neural networks (PNN) [9], [11]. The GRNN and PNN have a disadvantage: they perform the operations slower than other kinds of networks [12], [13].

B) Fuzzy Logic Techniques: Another interesting possibility, for clustering purposes, is the use of fuzzy methods: ANFIS [14], fuzzy C-means, originally introduced by Bezdek in 1981

TABLE I
MAIN FEATURES OF SUPERVISED AND UNSUPERVISED NEURAL NETWORKS

Kind of Network	Topology	Learning rule	Author/s
Simple perceptron	1 layer	Error correction	Rosenblatt (1958)
ADALINE	2 layers Feedforward	Error correction (Least mean square) Delta rule	Widrow/Hoff (1960)
Multilayer perceptrons (MLPs)	More than 1 layers Feedforward	Error correction	Rosenblatt
Learning vector quantization	2 layers feedforward	Competitive learning	Kohonen (1981)
Elman NN	2 layers feedback	Gradient of error	Elman (1990)
SOM	2 layers feedforward with lateral connection	Competitive learning (Kohonen learning rule)	Kohonen (1982)

TABLE II
CUSTOMER SPECTRUM

Customer description	Labels
Food Industry_A & B	1 & 2
Warehouse	3
Paper Industry_C & D	4 & 5
Plastic Industry_E & F	6 &7
Chemical Industry_G & H	8 & 9
Large University_1 & 2	10 & 11
University_Campus_1	12
University_Campus_2	13
Small Hotel	14
Medium Hotel	15
High School	16
Hospital and medical center_1	17
Hospital and medical center_2	18
Hospital and medical center_3	19
Hospital and medical center_4	20
Retailers_1 & 2	21 & 22
Residential C.T.	23

[15] as an improvement on earlier clustering methods, or substractive clustering [16].

Fuzzy methods entail work with data collections whose boundaries are not clearly defined by means of the so-called membership functions, which try to measure the affinity a sample of data has with respect to a cluster.

C) Statistical Techniques: Two main groups of techniques can be distinguished: multivariate statistics and intuitive techniques.

The first of them, multivariate statistics, includes those methods that consider a group of variables together rather than focusing on only one variable at a time to understand a data set.

Among all these techniques, MANOVA [17] has a special interest. MANOVA is a technique for assessing group differences across multiple metric-dependent variables, based on a set of categorical (non-metric) variables acting as independent variables. MANOVA uses one or more categorical independents as predictors, like ANOVA (analysis of variance), but unlike ANOVA, there is more than one interval dependent. Some MANOVA applications allow the following:

- — to compare groups formed by categorical independent variables on group differences in a set of interval dependent variables;
- — to use the lack of difference for a set of dependent variables as a criterion to reduce a set of independent variables to a smaller, more easily modeled number of variables;
- — to identify the independent variables that differentiate a set of dependent variables the most.

A second group of statistical techniques, the inductive ones, includes decision trees [18]. They have a great explanatory capacity but a poor predictive capacity, an interesting property in neural and fuzzy techniques.

B. Case Study

To evaluate the methodologies mentioned above, a set of measurements corresponding to a mix of industrial, institutional, commercial, and small residential loads (in this case, the load is aggregated at the high voltage side of a distribution transformer center, CT) has been used as input space. The annual load peak ranges from 100 kW to 10 MW.

The input database consists of 23 Spanish customers of the Mediterranean southeast coast. Energy data belong to January and February 2003/2004 months, and they correspond to weekly load curves. Vectors are normalized using the maximum month value of demand for each customer. Table II shows the description and the label associated to daily load curves for each customer (a label number for all the daily load curves).

C. Selection of Methodologies

Some algorithms were developed using Matlab toolbox libraries in order to evaluate the classification ability of the analyzed techniques (see Fig. 2). The objective was to select one of them in order to perform a more detailed study of customer clustering and identification features when several different treatments are applied to input data (customer demands).

The computer used was a Pentium IV CPU at 2.5 GHz and 512 Mb of RAM. Several tests with the customer profiles were made to check the clustering results for each methodology. Different training architecture configurations and training algorithms were used with each technique to find the best results.

Additionally, an index named "learning error" was defined to evaluate the relative quality of the learning and segmentation capacities and so to select the best configurations previously commented. This value represents the number of input demand vectors that the technique was not able to identify or classify correctly after the training step. Obviously, the same input data set was used for all the training sessions.

Most techniques were able to match correctly each vector with its correct label, but some of them were not able to match them all, some uncertainty appearing in the results. Table III shows the "learning error" for each one of the tested techniques.

Among statistical multivariate techniques, MANOVA is especially interesting: the technique allows us to see graphically the output data in a similar way to SOM maps so results can be easily analyzed. In Fig. 3, a multidimensional scaling (a non-dimensional representation showing relative distances between demand data) was performed to show the results of this method. Fig. 3 presents the high quality of clustering achieved through the MANOVA approach.

TABLE III
LEARNING ERROR

NEURAL NETWORK TECHNIQUES		
MLPs	Levenberg-Marquardt algorithm	4
	Resilient backpropagation	0
	Gradient descent backpropagation	123
Learning Vector quantization (LVQ)		5
ELMAN Neural Network		2
Self-Organizing Maps		0
Radial Basis Function Neural Network (RBF)		2
Generalized Regression Neural Network (GRNN)		1
Probabilistic Neural Network		1
FUZZY LOGIC TECHNIQUES		
ANFIS		1
Subtractive Clustering		2
STATISTICAL TECHNIQUES		
Discriminant Analysis		5
Decision Trees		2

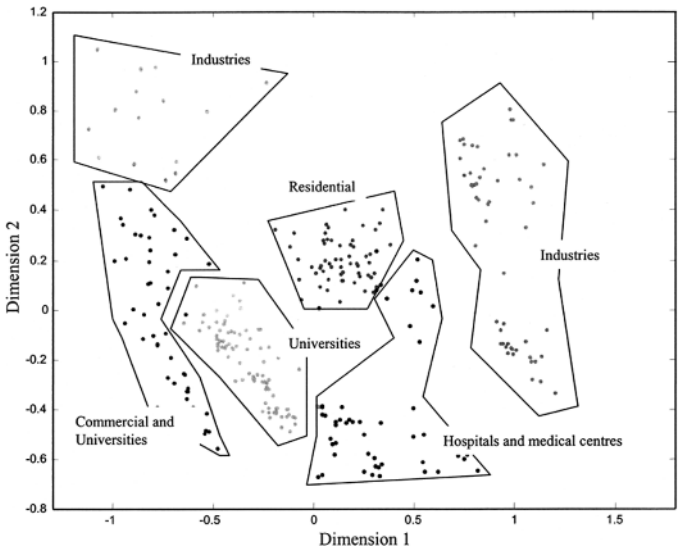

Fig. 3. Graph of clusters obtained with MANOVA and later multidimensional scaling.

Finally and regarding the results, two different groups of techniques were found:

— methodologies showing a considerable ability to classify and group the input space database, such as multidimensional scaling, fuzzy C-means clustering, MANOVA, and K-means clustering [19];

— methodologies showing an ability to classify the input space database and furthermore to identify new customer patterns when new customers or measurements increase the database (i.e., memory behavior). For example: MLP, RBF, GRNN, SOM, PNN, and ANFIS.

Some of these methodologies show good performances for the research interest: quick processing capacity, high quality results when the problem reaches high levels of complexity, and the ability to learn from a database to produce a further classification and identification when the input space growths.

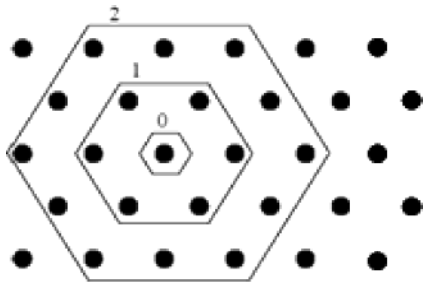

Fig. 4. Hexagonal grid.

Both MANOVA and SOM techniques are useful for customer clustering, but at this stage of the research work, a preference for SOM tools is reported due to the higher experience of the authors with this last technique and the availability of software tools.

In the following sections, some insights to improve SOM potential are presented.

III. SELF-ORGANIZING MAPS METHODOLOGY

This methodology was introduced by Kohonen two decades ago [6]. These networks are a kind of unsupervised ANN that performs a transform from the original input space (n dimensional data vector) to a reduced output space (bidimensional). The advantage of SOM is that the relationship between the original vectors is to some extent preserved in the output space, providing a visual format where a human operator can "easily" discover clusters, relations, and structures in the usually complex input space database.

The number of neurons can vary from a few dozen up to several thousands. Each neuron is represented by a d-dimensional weight vector (prototype vector, codebook vector) $\mathrm{m} = [\mathrm{m}_1, \ldots, \mathrm{m}_\mathrm{d}]$, where $\boldsymbol{d}$ is equal to the dimension of the input vectors. The neurons are connected to adjacent neurons by a neighborhood relation, which dictates the topology or structure of the map. This topology is defined by two factors: local lattice structure and global map shape. A hexagonal lattice structure and a sheet map shape were used (see Fig. 4). In this figure, discrete neighborhoods (size 0, 1, and 2) of the centermost unit are defined. The innermost polygon corresponds to the 0-neighborhood, the second to the 1-neighborhood, and the biggest to the 2-neighborhood.

The SOM training algorithm resembles vector quantization algorithms, such as K-means [19]. The important distinction is that, in addition to the best-matching weight vector, its topological neighbors on the map are also updated: the region around the best-matching vector is stretched toward the presented training sample, as in Fig. 5. The final result is that the neurons on the grid become ordered: neighboring neurons have similar weight vectors.

Since the weight vectors of the SOM have well-defined low-dimensional coordinates $\mathbf{r_i}$ on the map grid, the SOM is also a vector projection algorithm. Together, the prototype vectors and their projection define a low-dimensional map of the data manifold.

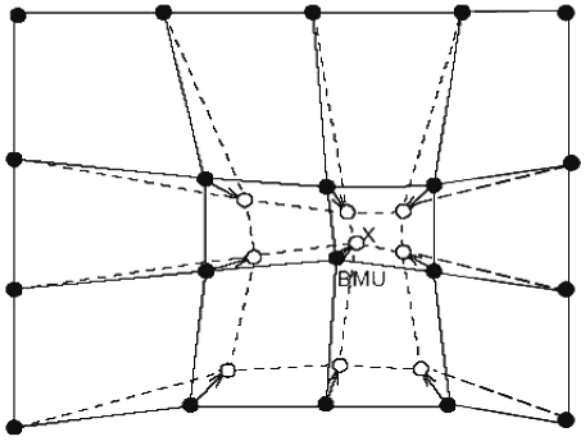

Fig. 5. Updating the best-matching unit (BMU) and its neighborhood. Toward the input sample marked with x. Solid and dashed lines correspond to situation before and after updating, respectively.

Along the research, different trainings of the maps with two algorithms were carried out: sequential training algorithm and batch training.

With the sequential training algorithm, the SOM is trained iteratively. In each training step, one sample vector $\mathbf{x}$ from the input data set is chosen randomly, and the distances between it and all the weight vectors of the SOM are calculated using some distance measures. The neuron whose weight vector is closest to the input vector $\mathbf{x}$ is called the best-matching unit (BMU), denoted here by $\mathbf{c}$

$$||x - m_c|| = \min_i \{||x - m_i||\} \tag{1}$$

where $||.||$ is the distance measure, typically a Euclidian one.

After finding the BMU, the weight vectors of the SOM are updated so that the BMU is moved closer to the input vector in the input space. This adaptation procedure stretches the BMU and its topological neighbours toward the sample vector as shown in Fig. 5.

The SOM update rule for the weight vector $\boldsymbol{m}$ of unit $\boldsymbol{i}$ is

$$m_i(t+1) = m_i(t) + \alpha(t) h_{ci}(t) \left[x(t) - m_i(t)\right] \tag{2}$$

where $\boldsymbol{t}$ denotes time, $\mathbf{x}(t)$ is an input vector randomly drawn from the input data set at time $\boldsymbol{t}$, $\boldsymbol{h_{ci}}(\mathrm{t})$ the neighborhood kernel around the winner unit $\boldsymbol{c}$, and $\boldsymbol{\alpha(t)}$ the learning rate at time $\boldsymbol{t}$. The neighborhood kernel defines the region of influence that the input sample has on the SOM.

The training is usually performed in two phases. In the first phase, relatively large initial learning rate $\boldsymbol{\alpha_o}$ and neighborhood radius $\boldsymbol{\sigma_o}$ are used. In the second phase, both learning rate and neighborhood radius are small right from the beginning.

Also the batch training algorithm is iterative, but instead of using a single data vector at a time, the whole data set is presented to the map before any adjustments are made (hence the name "batch"). In each training step, the data set is partitioned according to the Voronoi regions of the map weight vectors, i.e., each data vector belongs to the data set of the closest map unit. After this, the new weight vectors are calculated as follows:

$$m_i(t+1) = \frac{\sum_{j=1}^{n} h_{ic}(t) x_j}{\sum_{j=1}^{n} h_{ic}(t)} \tag{3}$$

where $\mathrm{c} = \arg\min_{\mathrm{k}}\{||\mathrm{x_j} - \mathrm{m_k}||\}$ is the index of the BMU of data sample $\boldsymbol{x_j}$. The new weight vector is a weighted average of the data samples, where the weight of each data sample is the neighborhood function value $\boldsymbol{h_{ci}(t)}$ at its BMU $\boldsymbol{c}$.

Notice that in the batch version of the K-means algorithm, the new weight vectors are simply averages of the Voronoi data sets. The above equation equals this if $h_{ci} = \delta(i, c)$. Alternatively, one can first calculate the sum of the vectors in each Voronoi set

$$s_i(t) = \sum_{j=1}^{nv_i} x_j \tag{4}$$

where $\boldsymbol{nv_i}$ is the number of samples in the Voronoi set of unit $\boldsymbol{i}$. Then, the new values of the weight vectors can be calculated as

$$m_i(t+1) = \frac{\sum_{j=1}^{m} h_{ij}(t) s_j(t)}{\sum_{j=1}^{m} nv_j h_{ij}(t)} \tag{5}$$

where $\boldsymbol{m}$ is the number of map units.

To summarize, in SOM methodology, the neurons become selectively tuned to various input patterns (stimuli) or classes of input patterns in the course of a competitive learning process. A SOM is therefore characterized by the formation of a topographic map of the input patterns in which the spatial locations (i.e., coordinates) of the neurons in the lattice are indicative of intrinsic statistical features contained in the input patterns, hence the name "self-organizing map."

IV. Application of SOM for Anomalous Behavior Filtering

The first task to accomplish the clustering process is to make a previous filtering of anomalous demand behaviors. To analyze the possibilities of SOM for load data filtering, a university was selected from the customer case study. Obviously, these records (196) include some anomalous days and wrong measurements.

An alternative labeling to the one proposed in Section II is used for a better understanding of results. By means of this labeling, a number is assigned to each load profile following the next criterion: the last two digits indicate the day of the month and the initial remaining ones the corresponding month (mm/dd). Thus, a label map (see the upper part of Fig. 6) allows the identification of daily load data assigned to each cell.

The information contained in the daily load curves is directly presented to the map, allowing a fast input from database records (a detailed discussion about the input data format is analyzed in Section V). Specifically, load demand curves were used per unit, recorded every 15 min. The reason was the good results obtained in previous works [20], [21], some of them accomplished by the authors [22].

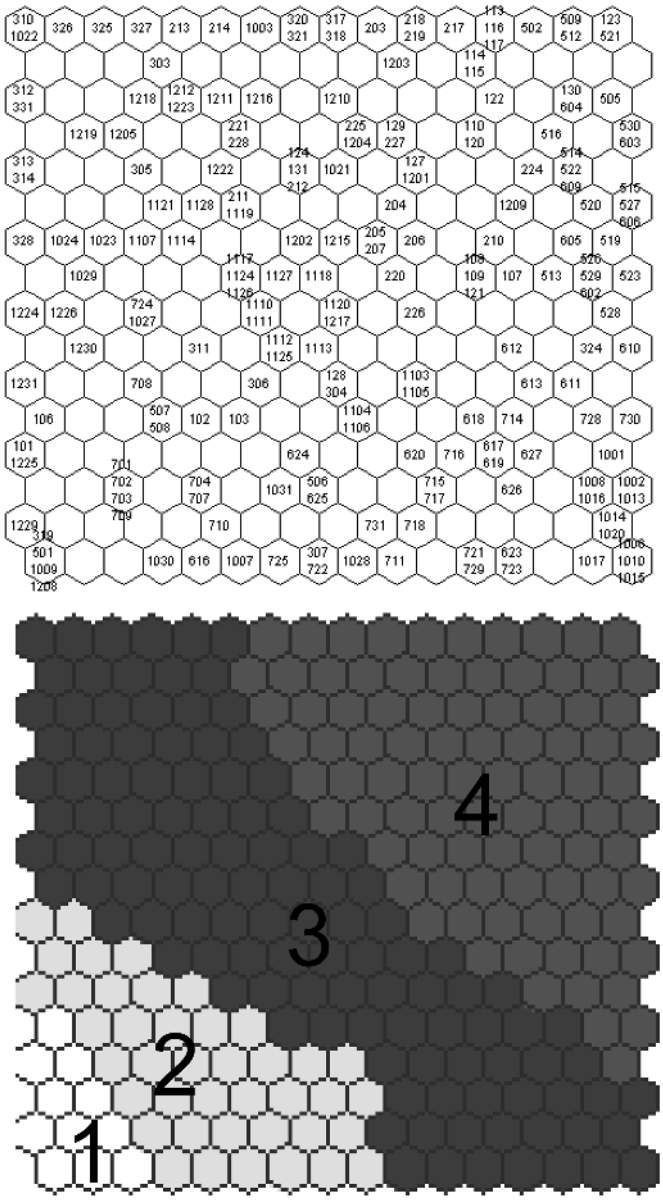

Fig. 6. (Upper) University label (mm/dd criterion). (Lower) Cluster maps.

A hexagonal network formed by a total of 256 neurons (16×16) was used. This size has been chosen to allow a better visualization of the output data of the training map. A network with a greater number of cells would have hindered the visualization of the labels in each neuron. In the same way, a smaller map than the one used by the authors would cause many labels to be overlapped.

Finally, random initialization of the map and a batch training algorithm with 1000 and 500 steps for the rough and the fine-tuning training, respectively, was used.

The minimum number of epochs for "rough" phase and for "finetune" phase to achieve a correct network convergence in the training are, respectively, $10 \times \mathrm{Mpd}$ and $40 \times \mathrm{Mpd}$, where Mpd is the ratio of number of neurons of the map to the number of data samples.

For longer training times than mentioned above, a correct map convergence is always achieved after the training.

Once the network is trained, it is possible to force data clustering on the map presented in the upper part of Fig. 6. After some tests, the four zones defined in the lower part of Fig. 6 were found.

For example, the upper part of Fig. 6 shows how labels 501 (May 1) and 1208 (December 8), corresponding to holidays in Spain, were both located on the left bottom area of the map. Also a county holiday marked with label 1009 (October 9) is located close to the previous ones. Besides, two cells at the border of regions 1 and 2 (left bottom are in the lower part of Fig. 6: cells 1231, 106) correspond to holidays in the Christmas season.

By means of the label map and plotting the corresponding load profiles, it can be seen that the network is able to distinguish three kinds of load profiles: typical consumption patterns, assigned to regions 3 and 4; holiday profiles (placed in region 1 due to the SOM characteristic of topographic preservation); and finally, profiles that denote a different behavior from the usual one located in region 2 (July days when students take their exams and the building occupation is lower). Besides, the filtering process presents other applications: the detection of erroneous measures (failure in demand meters) and particular behaviors of the customer (low demand periods due to holidays). This last characteristic reduces the possibility of clustering failures (for example, university holiday demand is near to the typical demand profile of some industry customers).

V. Customer Patterns Classification

It has been stated in previous paragraphs that SOM map is a valuable tool to group (aggregate) and classify (disaggregate) electrical customer patterns. This section explains how to improve a classification tool such as SOM maps through the analysis of the influence of the form of the M data set arrays $(1 \times \mathrm{N})$ used to feed and train the map.

Thus, each M data set array reflects the load behavior associated to an elemental user demand included in the customer case study (see Section II). It should contain the necessary information to evaluate the affiliation of each elemental demand to a cluster. Traditionally, this customer clustering was based on the type of economical activity declared by users (for example, through NACE codes [23]) and voltage levels, but this approach has not proven to be as efficient as is possible because several patterns can be found for the same economical activity, or users with different activity can show similar demand patterns.

From the point of view of the authors and technical interests (demand response and distributed generation), it is necessary to find similar load characteristics, and this can be reached through field measurements performed by the customer or by commercializers to obtain, reduce, and manage energy and power costs. Standard measurement devices in Spain usually have a pacer trigger of four samples an hour.

These M data set arrays (power versus time records) are the input space in [21] and [24], where the major improvement in customer classification is focused on the ANN used for load pattern recognition. Besides, in [21], the measured load demand refers to working days, i.e., weekend demand profiles are not considered. This hypothesis does not make much sense when the objective is to develop dedicated tariffs rates [24], and weekend demand can have an important influence in its

design (for example, some industries, hospitals, and hotels that usually work the whole weekend). The apparent justification of this approach is the growth in size of each input data set vector $\mathrm{N} \times (1 \times 96 \text{ samples/day})$, where N is seven days, that perhaps makes more difficult the performance of SOMs. A way to include this relevant information while reducing input vector sizes in a clever way is established in the next sections.

A. Transform of Demand Data From Time Domain to Frequency Domain

The idea was to extract as much information as possible while compressing, filtering, and simplifying the available information (weekly demand). Perhaps a simpler input array would include all the relevant information about customer demand behavior and also would improve the topological projection of SOM maps, i.e., a double transform from time domain to SOM output domain will be tried through a frequency transform in order to obtain some improvements on customer clustering.

Several approaches are available to compress and transform information from time domain to signal frequency domain. This problem is broadly used to solve other problems in power systems, such as load forecasting. For example, in [25], a Fourier series analysis is applied to filter load data before an ARIMA model is applied. In [26], a wavelet transform is also proposed to obtain a short-time load model. This last transform has been broadly used to extract anomalous patterns in the transient analysis of power systems. From this knowledge of main applications of the wavelet transform, the fast Fourier transform (FFT) was selected as the most interesting transform to extract steady-state demand behavior from demand profile records.

FFT performs the discrete Fourier transform of a certain waveform and allows us to find the more representative harmonics. This means that it can be easily observed if a certain behavior in a specified time period (day, week, etc.) appears. As it was previously stated, the objective is to find the load behavior in a day or in a week, including the weekend. For this reason, individual demand curves were treated in the following way:

- — extract daily load curves: the process of obtaining (1×96) vectors;
- — filtering of anomalous data vector (see Section IV);
- — extract working days (WD_vectors): to select labor days in each week;
- — aggregate working days in a new vector (WDS_vectors): from Monday to Friday if anomalous days are not found;
- — extract week vectors (WES_vectors): vectors of seven days.

Applying the FFT function to each time-domain vector (WD, WDS, and WES), and the corresponding equations to obtain Fourier Series coefficients (sinusoidal form), the following vectors were obtained.

- — FWD_vector: mean value $(\mathrm{n} = 0)$ and sine and cosine harmonic terms $(\mathrm{n} = 1, 2 \ldots)$. Thus, the fundamental $(\mathrm{n} = 1)$ shows sine and cosine terms with period $\mathrm{T} = 1$ day (frequency $= 1/\text{day}$).
- — FWDS_vector: mean value $(\mathrm{n} = 0)$ and sine and cosine harmonic terms $(\mathrm{n} = 1, 2 \ldots)$. The first term $(\mathrm{n} = 0)$ is the average demand in working days and the fundamental $(\mathrm{n} = 1)$ shows sine and cosine terms with period $\mathrm{T} = 5$ days.
- — FWES_vector: mean value $(\mathrm{n} = 0$; average demand in a week), the fundamental $(\mathrm{n} = 1)$ shows the sine and cosine terms with period $\mathrm{T} = 7$ days (frequency $= 1/\text{week}$).

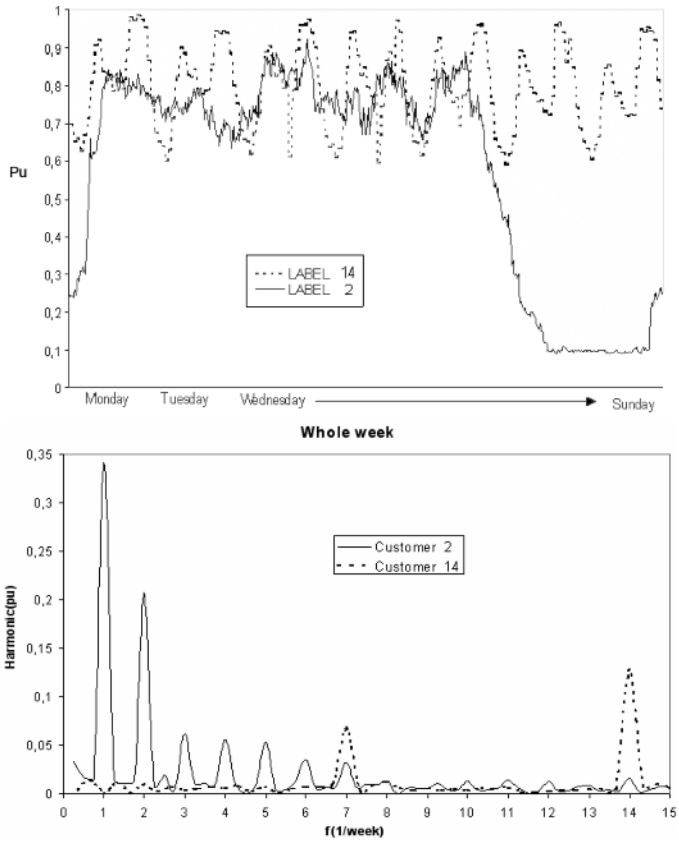

Fig. 7. (Upper) Week load profiles and (lower) FWES_vector for customer 2 and 14.

Thus, each input vector presented to SOM map will have a mix of frequency treatments extracting frequency-domain information from some representative terms of FWD, FWDS, and FWES vectors. After some tests performed with several selections and weighting of harmonics terms, a combined frequency-domain FD_vector was chosen defined as

$$\begin{aligned}\text{FD\_vector} = &\{\text{FWD\_vector(i)};\ \mathrm{i} = 0 \ldots 12\}\ \& \\ &\{\text{FWDS\_vector(j)};\ \mathrm{j} = 5, 10\}\ \& \\ &\{\text{FWES\_vector}\ (\mathrm{k});\ \mathrm{k} = 0, 1, 7, 14\}\,.\end{aligned}$$

In this way, the daily load demand is transformed into the average and the 12 first daily sine and cosine harmonic terms (a total of $1 + 24$ terms) plus two terms from FWDS vector ($\mathrm{n} = 5$ and $\mathrm{n} = 10$; four terms) accounting for daily pattern "filtering" in labor days. Finally, some terms from FWES (mean demand, and $\mathrm{n} = 1$, 7 and 14 sine and cosine terms; i.e., 7 terms) were added to force the SOM network to account the weekend load behavior. Obviously, last 11 terms of FD_array are the same for each day in a week.

For a better understanding of this procedure, Fig. 7 shows the weekly demand profile and its transform (FWES) for two

customers (labels 2 and 14). The lower part of Fig. 7 justifies the harmonic terms of FWES. The customer 14 shows a weekend demand behavior similar to working days, so the harmonic array FWES shows a low $\mathrm{n} = 1$ term. Notice the value of harmonic $\mathrm{n} = 14$ due to the daily demand fluctuation (two peaks a day). However, the customer 2 exhibits a weekend load reduction, and terms $\mathrm{n} = 1, 2$ are quite a lot higher. The term $\mathrm{n} = 0$ has not been presented for simplicity.

VI. Results: SOM Clustering and Identification Performance

A. Customer Clustering

Different policies have been selected to feed a SOM network and thus to test the usefulness of the Fourier transform. Two cases were evaluated.

- — Time-domain case: the input domain is a set of 1×96 data vectors corresponding to working-day demand versus time profiles, i.e., the approach presented in [21].
- — Frequency-domain case: input arrays are a set of 1×36 data vectors in the way it was presented in paragraph V (FD_vector).

To carry out the projection from the original data set space (374 filtered working days) to the SOM output space, the use of a $\mathrm{N} \times \mathrm{N}$ lattice is proposed. The choice of N is subjective; some authors suggest the use of a number of map cells lower than the number of samples [21]. In this case and in order to promote a better visual definition of clusters, a 20×20 size map, slightly higher than the number of samples $(374 > 19 \times 19 = \mathrm{N} \times \mathrm{N})$, was selected.

For each case, different training possibilities arise: linear, sequential, and random trainings were tested. Besides, different combinations were applied and the number of 5000 and 3000 steps, for primary and secondary training, respectively, was finally applied.

For a better understanding of SOM maps, it is important to note that different training (randomly) sessions usually produce a different map even for the same data set. Notice that these maps conserve the relative position between the elemental cells but not their absolute position for different training sessions. Also, the time needed for training each map is quite different due to the complexity of each input data set. In this case, the time ranges from 20 min for time-domain training to 13 min for frequency-domain training.

Finally, the selection of the number of clusters is another significant task. This number, a subjective value, should be a reasonable one between two obvious options: the number of macro-clusters in the customer case study (residential, commercial, industrial) and the overall number of customers. In this way, the number of clusters should allow an average customer aggregation of more than a customer per cluster. For this reason, an automatic selection of an optimal number of clusters is found after applying K-means function for 12 as the maximum number of clusters (23 customers). An optimum clustering can be guaranteed by the minimization of Davies–Bouldin (DB) index (a detailed explanation of this index can be found in [27]), but in some cases, visual inspection helps the researcher to decide the clustering.

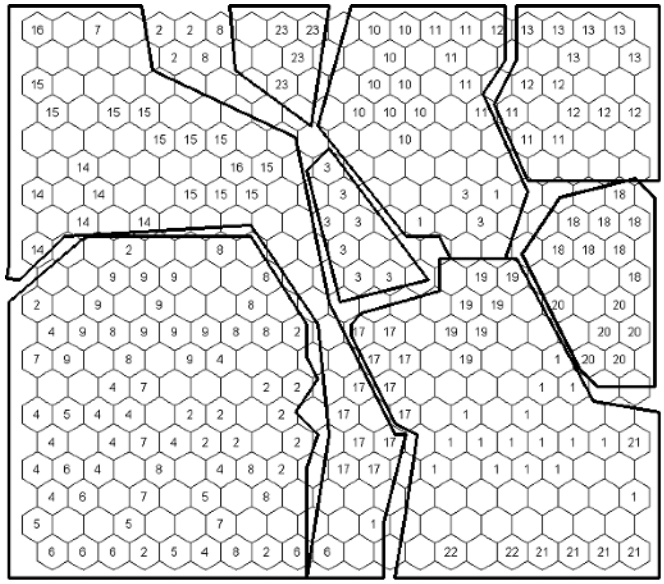

Fig. 8. SOM map training with time-domain values.

Small values of DB index correspond to good clustering results (the clusters are compact, and their centers are far away from each other). The cluster configuration that minimizes DB is taken as the optimal number of clusters. The results are shown in the following paragraphs.

1) Time-Domain Approach: In this case, 374 daily load profiles sampled every 15 min have been presented to the SOM map (see Fig. 8). The map shows the aggregation of labels and the clusters found after applying the K-means function. In this case, DB index reached a minimum value (0.82) for a number of eight clusters.

Several conclusions can be inferred: the aggregation process is quite good; only two single-customer clusters appear (labels 3 and 23). Universities are split into two clusters, and main industrial customers are grouped in a big cluster (except labels 1 and 3). However, the map has some problems, too: customers 11 and 17 are classified in two different clusters, and besides, some cells (2 and 8 in the upper left side of the map) are not assigned to a specific cluster.

2) Frequency-Domain Approach: Vectors "FD" with daily and weekend harmonic values have been used for 20×20 SOM map training. The aggregation of labels and the clusters found after applying the K-means function are shown in the left part of Fig. 9. In this case, the DB index has a minimum value (1.19) for a number of seven clusters. A sparsely filled map is the main characteristic of this approach. The labels are closer, and there is not any cluster error in the label location process. Notice that a cell in the map can often contain several customer profiles from the input space, but only the most repeated label is shown for a better understanding (see zoom in Fig. 9). Industries are split in several clusters: customers with high weekend demand (labels 4 to 7) and industrial customers without continuous demand during the weekend (labels 2, 8, and 9; see the right part of Fig. 9). Other clusters are: universities (10 to 13), retails (20 and 21), medical centers (18 and 20), residential cluster (23), and two clusters whose customers have different activities (14

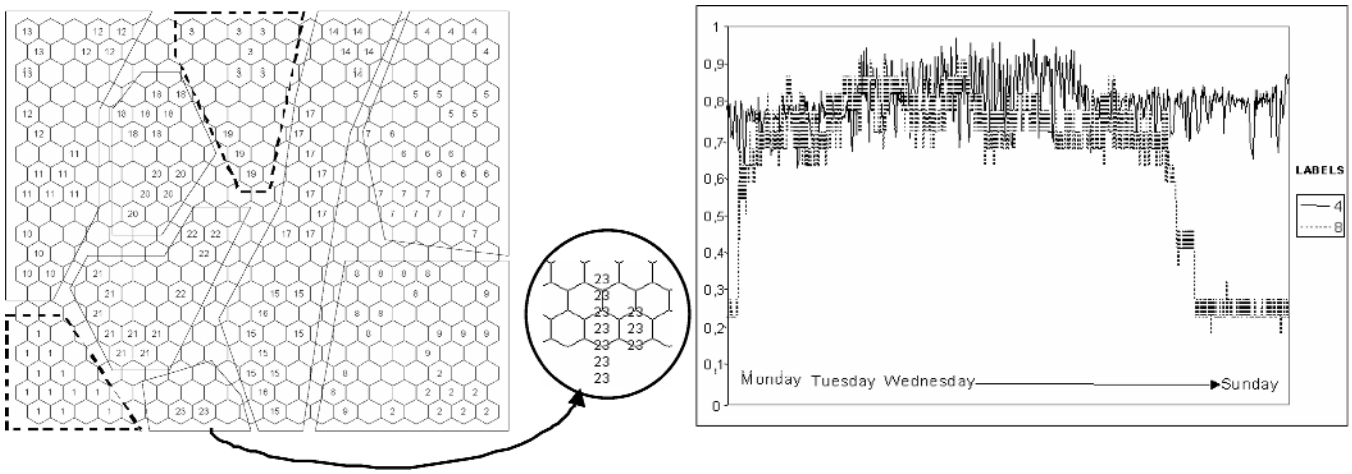

Fig. 9. (Left) SOM training with frequency-domain and (right) corresponding weekly load profiles (customers 4 and 8).

to 17 and 1, 3, and 19, the cluster with a dashed line) but similar demand behavior.

At first sight, the map seems to get a better customer clustering, but it is necessary to measure the map's quality in terms of some analytic indexes. The issue of SOM quality is a complicated one. Typically, two evaluation criteria are used: resolution and topology preservation. If the dimension of the data set is higher than the dimension of the map grid, these usually become contradictory goals. This quality is analyzed in terms of mean quantization error $(\mathrm{Q_e})$, which measures the resolution of the map, and the topographic error $(\mathrm{T_e})$ [28], which measures the distortion of the map. $\mathrm{Q_e}$, also known as average quantization error, is simply the average distance (weighted with the mask) from each data vector to its BMU [28]. The topographic error is the proportion of all data vectors for which first and second BMUs are not adjacent units. During the training of SOMs, there was a conflict between the twin goals of topology preservation, between input and output and the minimization of quantization error $(\mathrm{Q_e})$. This is especially obvious when the dimension of the input data (the dimension of the codebook vectors) is higher than the dimension of the output network (the dimension of the map grid). The average quantization error is calculated over the input samples, and it is defined as

$$Q_e = \frac{1}{N}\sum_{i=1}^{N} ||x_i - m_c(x_i)|| \tag{6}$$

where "N" is the number of input vectors of the data set, "$\boldsymbol{x}_i$" is each input vector, "$\boldsymbol{m}$" is the weight vector, and "$\boldsymbol{c}$" indicates the BMU for "$\boldsymbol{x}$." After training, for each input sample vector, the BMU in the map is searched for, and the average of the respective quantization errors is returned.

A simple method for calculating the topographic error

$$T_e = \frac{1}{N}\sum_{k=1}^{N} u(x_k) \tag{7}$$

where $\boldsymbol{u}(\boldsymbol{x_k})$ is 1 if the first and second BMUs of $\boldsymbol{x_k}$ are not next to each other. Otherwise, $\boldsymbol{u}(\boldsymbol{x_k})$ is 0.

TABLE IV
SOM QUALITY ANALYSIS

Training Case	$\mathrm{T_e}$	$\mathrm{Q_e}$
Time-domain	0.008	0.3065
Frequency-domain	0.003	0.041

The net advantage of frequency-domain transform is well established from these indexes (see Table IV).

B. Identification of New Customers

The second objective is to show the capacity of SOM for customer classification. Two new customers, unknown by the SOM network (a mall and a restaurant), were used to test the SOM adequacy for new customer classification. Again, time and frequency approaches were evaluated. The target was to get for each new input data set the most similar cell.

Two validation tests were developed: a visual test (see Fig. 10) and an analytic one based on quantization error (see Table V). Both methods are based on BMU function supplied with SOM toolbox. This function supplies the cell or neuron (and label if available), in a previously trained SOM map that is close to each new input vector, and its corresponding quantization error $(\mathrm{Q_e})$.

The first map, trained in the time domain, was able to classify without problems the set of daily load curves corresponding to the first new customer (mall). In this case, the user is located near the "university cluster" (labels 12 and 13 in the left part of Fig. 10).

The second customer (restaurant) presented a greater uncertainty, and the SOM does not present a clear result (up to three clusters were related to the new customer; labels 3, 15, and 23 in the left part of Fig. 10).

The second map, trained in the frequency domain, shows the best result. New customers (mall and restaurant) are located in a unique cluster ("residential cluster" for the restaurant and "campus university cluster" for the mall. See labels 23, 12, and 13 in the right part of Fig. 10).

Table V verifies analytically the conclusions stated in the previous paragraphs. The SOM map trained in the time domain shows higher values of the $\mathrm{Q_e}$ index (a worse identification of

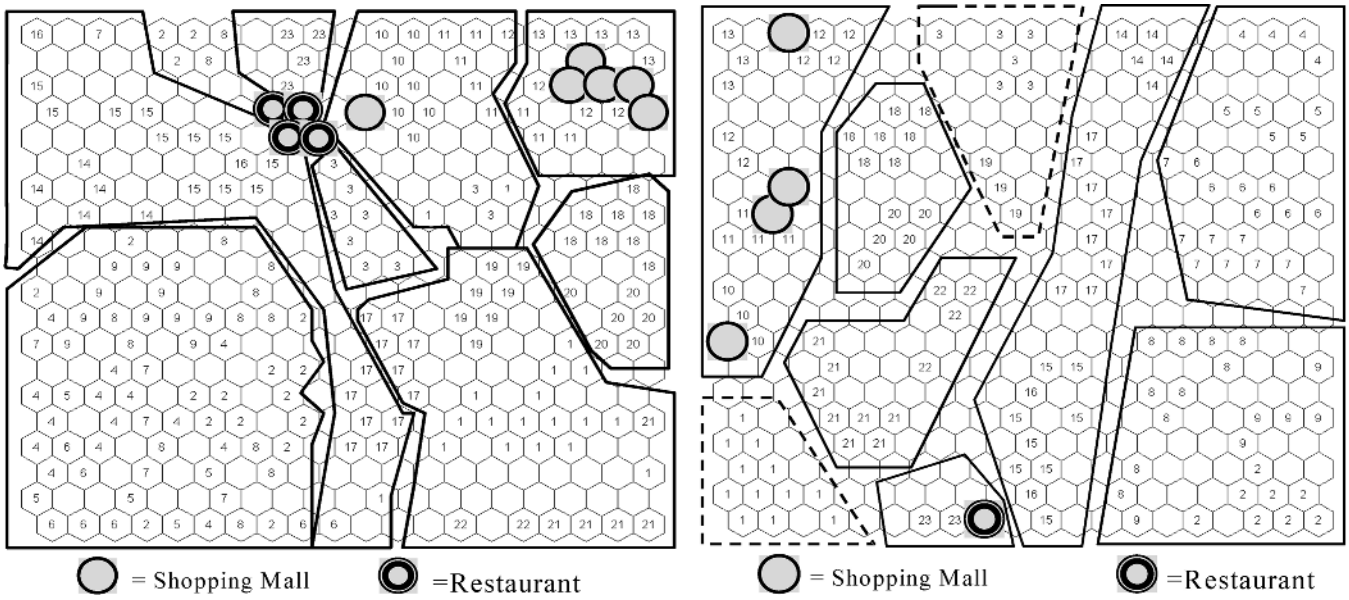

Fig. 10. (Left and right) Classification of new customers.

TABLE V
RESULTS OF CUSTOMER TESTING

SOM maps	Customer tested	Nearest label identified	Identified clusters	Q_e (mean)
Time-domain	Mall	1 cell with label 10 3 cells with label 13 15 cells with label 12 11 cells between 12 &13	Large Univ. (3%) Campus Univ. (97%)	1.24
	Restaurant	7 cells near label 23 2 cells near label 23 9 cells between 15&23 17 cells near label 3	Residential (39%) Hotels (13%) Warehouse (48%)	0.93
Frequency-domain	Mall	23 cells with label 10 2 cell with label 11 5 cells near label 12	University (100%)	0.28
	Restaurant	35 cells near label 23	Residential (100%)	0.21

the new customer) for both customers: mall ($Q_e = 1.24$) and restaurant ($Q_e = 0.93$). On the contrary, SOM map trained in the frequency domain presents the minimum values for the Q_e index: mall ($Q_e = 0.28$) and restaurant ($Q_e = 0.21$).

From these results (see Fig. 10 and Table V), it can be concluded that the classification of new users is more accurate when a SOM frequency-domain map is used.

VII. CONCLUSIONS

A SOM development is presented to achieve the segmentation and demand patterns classification for electrical customers on the basis of database measurements. In case of presence of anomalous data, some uncertainty appears. An ANN tool also provides the effective detection of outliers from standard pattern, due to external factors, as it is the case of external temperature growth.

The frequency transform proposed in this paper to extract information from original demand profiles shows an improvement in clustering performance (see Q_e and T_e indexes reduction) and a better accuracy in new customer classification. Notice the significance of some harmonics in weekend pattern recognition (influence not considered in previous works) and the compression rate of input data in comparison to the original customer case study.

The method presented here can effectively help commercializers and distributors in customer segmentation and classification. This is the first step to evaluate cost-effectiveness of a lot of necessary policies in the demand-side: the potential of energy efficient alternatives, customer response to real price or TOU tariffs, the success of dual-fuel or energy storage appliances, or the possibilities of distributed generation in medium and small users. The future research activity, already under study, is devoted to the development of three objectives: the improvement of segmentation indexes used in the SOM map, the study of the potential and applicability of other promising clustering techniques (see Section II and specifically MANOVA), and the development of new tools based on ANN to identify the potential interest of some customers to participate in short-term electricity markets. The results of these works will be reported in the future.

REFERENCES

[1] C. Alvarez, R. P. Malhamé, and A. Gabaldón, "A class of models for load management and application and evaluation revisited," *IEEE Trans. Power Syst.*, vol. 7, no. 4, pp. 1435–1443, Nov. 1992.

[2] C. Alvarez, A. Gabaldón, and A. Molina, "Assessment and simulation of the responsive demand potential in end-user facilities," *IEEE Trans. Power Syst.*, vol. 19, no. 2, pp. 1223–1231, May 2004.

[3] C. Olaru and L. Wehenkel, "Data mining tutorial," *IEEE Comput. Appl. Power*, vol. 12, no. 3, pp. 19–25, Jul. 1999.

[4] B. D. Pitt and D. S. Kirschen, "Application of data mining techniques to load profiling," in *Proc. IEEE PICA*, Santa Clara, CA, May 16–21, 1999, pp. 131–136.

[5] F. Rosenblatt, *Principles of Neurodynamics*. Washington, D.C.: Spartan, 1961.

[6] T. Kohonen, *Self-Organisation and Associative Memory*, 3rd ed. Berlin, Germany: Springer-Verlag, 1989.

[7] J. L. Elman, "Finding structure in time," *Cogn. Sci.*, vol. 14, pp. 179–211, 1990.

[8] C. Lau, Ed., *Neural Networks. Theoretical Foundations and Analysis*. Piscataway, NJ: IEEE Press, 1991.
[9] S. Chen, C. F. N. Cowan, and P. M. Grant, "Orthogonal least squares learning algorithm for radial basis function networks," *IEEE Trans. Neural Netw.*, vol. 2, no. 2, pp. 302–309, Mar. 1991.
[10] P. D. Wasserman, *Advanced Methods in Neural Computing*. New York: Van Nostrand Reinhold, 1993, pp. 155–161, and pp. 35–55.
[11] D. Gerbec and S. Gasperic *et al.*, "Allocation of the load profiles to consumers using probabilistic neural networks," *IEEE Trans. Power Syst.*, vol. 20, no. 2, pp. 548–555, May 2005.
[12] D. F. Specht, "Probabilistic neural networks and the polynomial adaline as complementary techniques for classification," *IEEE Trans. Neural Netw.*, vol. 1, no. 1, pp. 111–121, Mar. 1990.
[13] G. Chicco, R. Napoli, and F. Piglione, "Comparison among clustering techniques for electricity customer classification," *IEEE Trans. Power Syst.*, vol. 21, no. 2, pp. 933–940, May 2006.
[14] J.-S. R. Jang, "ANFIS: Adaptive-network-based fuzzy inference system," *IEEE Trans. Syst., Man, Cybern.*, vol. 23, no. 2, pp. 665–685, May-Jun. 1993.
[15] J. C. Bezdec, *Pattern Recognition With Fuzzy Objective Function Algorithms*. New York: Plenum, 1981.
[16] B. S. Suryavanshi, N. Shiri, and S. P. Mudur, "An efficient technique for mining usage profiles using relational fuzzy subtractive clustering," in *Proc. Int. Workshop Challenges Web Information Retrieval Integration*, Apr. 8–9, 2005, pp. 23–29.
[17] R. J. Harris, "Multivariate analysis of variance," in *Applied Analysis of Variance in Behavioral Science. Statistics: Textbooks and Monographs*, L. K. Edwards, Ed. New York: Marcel Dekker, 1993, vol. 137, pp. 255–296.
[18] H. Andrade, T. Kurc, A. Sussman, and J. Saltz, Decision tree construction for data mining on clusters of shared-memory multiprocessors Tech. Rep. CS-TR-4203 and UMIACS-TR.
[19] J. Hartigan and M. Wong, "A k-means clustering algorithm," *Appl. Stat.*, vol. 28, no. 1, pp. 100–108, 1979.
[20] A. Nazarko and Z. Styczynski, "Application of statistical and neural approaches to the daily load profiles modeling in power distribution systems," in *Proc. IEEE Transmission Distribution Conf.*, New Orleans, LA, 1999, vol. 1, pp. 320–325.
[21] G. Chicco, R. Napoli, F. Piglione, P. Postolache, M. Scutariu, and C. Toader, "Load pattern-based classification of electricity customers," *IEEE Trans. Power Syst.*, vol. 19, no. 2, pp. 1232–1238, May 2004.
[22] S. Valero *et al.*, "Characterization and identification of electrical customer through the use of SOM and daily load parameters," in *Proc. IEEE PSCE*, New York, Oct. 10–13, 2004.
[23] List of Nace Codes. Gateway to the EU [Online]. Available: http://europa.eu.in/comm/competition/mergers/cases/index/nace_all.html.
[24] G. Chicco, R. Napoli, P. Postolache, M. Scutariu, and C. Toader, "Customer characterization options for improving the tariff offer," *IEEE Trans. Power Syst.*, vol. 18, no. 1, pp. 381–387, Feb. 2003.
[25] F. D. Galiana *et al.*, "Identification of stochastic electric load models from physical data," *IEEE Tran. Autom. Control*, vol. AC-19, pp. 887–893, 1974.
[26] C. M. Huang and H. T. Yang, "Evolving wavelet-based networks for short-term load forecasting," *Proc. Inst. Elect. Eng., Gen., Transm., Distrib.*, vol. 148, no. 3, pp. 222–228, May 2001.
[27] D. L. Davies *et al.*, "A cluster separation measure," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-1, pp. 224–227, Apr. 1979.
[28] Helsinki University of Technology, SOM Toolbox for Matlab 5.0. [Online]. Available: http://www.cis.hut.fi/projects/somtoolbox/download/.

Sergio Valero Verdú was born in Elche, Spain, in 1974. He received a degree in industrial engineering in 1998 from the Universidad Politécnica de Valencia, Valencia, Spain.

Currently, he is an Associated Professor at the Universidad Miguel Hernández de Elche, Elche, Spain. His research activities include distribution system analysis, electricity markets, distributed energy resources, demand-side bidding, and neural network applications in power systems.

Mario Ortiz García was born in Murcia, Spain, in 1978. He received a degree in industrial engineering in 2002 from the Universidad Politécnica de Cartagena, Cartagena, Spain.

Currently, he is an Associated Professor at the Universidad Miguel Hernández de Elche, Elche, Spain. His research activities include wavelet and Hilbert applications to electricity, distribution system analysis, electricity markets, distributed and renewable energy resources, and neural network applications in power systems.

Carolina Senabre received a degree in engineering in 1998 from the Universidad Politécnica de Valencia, Valencia, Spain. She is currently pursuing the Ph.D. degree at the Universidad Miguel Hernández de Elche, Elche, Spain.

In 2001, she became an Associated Professor of mechanical engineering at the Universidad Miguel Hernández and has collaborated in several projects within the electrical engineering area regarding the electricity markets. She has authored numerous publications and contributions to congresses.

Antonio Gabaldón Marín (M'96) was born in Cieza, Spain, in 1964. He received the industrial engineering and Ph.D. degrees from the Universidad Politécnica de Valencia, Valencia, Spain, in 1988 and 1991, respectively.

Currently, he is a Full Professor at the Universidad Politécnica de Cartagena, Cartagena, Spain. His research activities include distribution system analysis, electricity markets, demand modeling and aggregation, distributed energy resources, energy efficiency, and demand-side management and response.

Francisco J. García Franco was born in Cartagena, Spain, in 1979. He received the industrial engineering degree in electrical power systems in 2003 from the Universidad Politécnica de Cartagena. Currently, he is pursuing the Ph.D. degree at the Institute of Energy Engineering, Universidad Politécnica de Valencia, Valencia, Spain.

His research activities include electricity markets, demand modeling and aggregation, demand-side bidding, and electrical customer classification.