International Journal of Power and Energy Systems, Vol. 28, No. 1, 2008

APPLICATION OF SELF-ORGANIZING MAPS FOR CLASSIFICATION AND FILTERING OF ELECTRICAL CUSTOMER LOAD PATTERNS

S.V. Verdú,* M.O. García,* C.S. Blanes,* F.J.G. Franco,** and A.G. Marín***

Abstract

The objective of this research is to show the capability of the self-organizing maps (SOMs) to organize, filter, classify and extract patterns from distributor, commercializer, aggregator or customer electrical demand databases (the objective known as data mining). This approach basically uses – to reach the above-mentioned objectives – the historic load demand curves of each user. To get a better classification, some anomalous data – holidays, wrong measurements due to recorder failures – should be filtered before starting the map training. This preliminary step has been performed through an SOM map too. To show the proposed method in the paper only two typical medium users are studied on the filtering stage: an industry and a university both located in Spain. Subsequently, the filtering process is applied to a larger group of customers to finally prove the customer clustering capacity of SOM. The results clearly show the suitability of SOM approach to improve data management and to find easily coherent clusters between electrical users.

Key Words

Demand management, self-organizing maps, electrical customer segmentation, load patterns

1. Introduction

The deregulation process began in developed countries a decade ago trained by political and technological reasons. Unfortunately, the experience has not been as successful as was planned, due to a lot of problems which have appeared since 2000 up to now, e.g., California Energy Crisis in 2000 or blackouts in Europe, 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 takes into account not only supply resources – Distributed Generation – but also some demand policies, e.g., efficiency gains in demand – in long term horizon – or price responsiveness – in short-term horizon. This supposes a new scenario where demand and supply compete on an equal footing in energy markets. For example, California Energy Commission will finance new energy efficiency programs to achieve a forecasted demand reduction of 6,000 GWh in 2008 [1]. The effective contribution to these energy efficiency programs and the necessity of offering energy choices to consumers need a detailed knowledge of customer segments and the characterization of these clusters, from the point of view of energy uses.

Besides, this new regulated framework of electrical power systems has promoted the necessity of new customer – and system – measurement, monitoring and control activities. This fact has increased the amount of data stored by supply side actors. The 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 modelled and advanced tools for synthesizing structures from such information are needed.

This is the main objective of the so-called data mining techniques (DM) or more precisely knowledge discovery in databases (KDD) [2, 3], and it is also the research purpose of this work applied to customer characterization and segmentation.

2. Review of Customer Classification Methodologies

In previous research studies [3–5], the clustering ability, for data segmentation and aggregation, of different techniques has been compared. The techniques used in this paper in an initial approach can be grouped in two categories: neural networks and fuzzy logic techniques.

Work on artificial neural networks (ANNs) has been motivated right from its inception by the recognition that the human brain computes in an entirely different way from

* Universidad Miguel Hernández Avda. de la Universidad s/n, 03202. Elche, Spain; e-mail: {svalero, mortiz, csenabre}@umh.es
** Institute of Energy Engineering (IEE), U. Politécnica de Valencia, Spain; e-mail: garfrafr@iie.upv.es
*** Department of Electrical Engineering, Universidad Politécnica de Cartagena, Spain; e-mail: antonio.gabaldon@upct.es
Recommended by Dr. Carlos Maria Alvarez
(paper no. 203-3825)

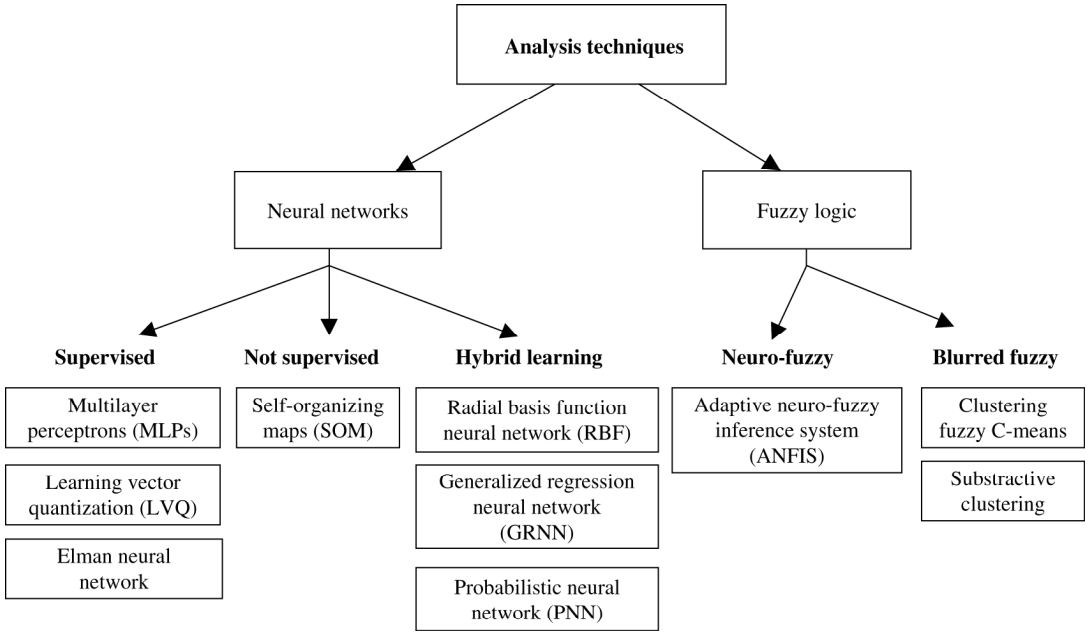


Figure 1. Classification methodologies review.

the conventional digital computer. The brain is a highly complex, nonlinear, and parallel information processor. It has the capability to organize its structural constituents, known as neurons, so as to perform certain computations many times faster than the most powerful digital computer in existence today.

Neural networks, or artificial neural networks, in a more accurate way, represent a technology that is rooted in many disciplines: neurosciences, mathematics, statistics, physics, computer science, and engineering. Neural networks find applications in such diverse fields as modelling, time series analysis, pattern recognition, signal processing, and control by virtue of an important property: the ability to learn from input data with or without a teacher. Fig. 1 shows a compendium of techniques used as clustering tools.

3. Selection of Methodologies

With the help of a personal computer several scripts were developed using a Matlab® toolboxes to evaluate the classification ability of the proposed techniques and select one of them to perform later a more detailed study of customer clustering and identification features when the input space representation changes (filtering, data compression, possibilities of mathematical transforms, etc.).

The results with the methodologies presented in Fig. 1 are subsequently reviewed. Simple Perceptron [6] and Adaptive Linear Element Adaline – are the first approach in ANN techniques. Only preliminary experiences with two kinds of customers were accomplished successfully because they do not work as well when they have to treat with large groups of customers due to their own limitations such as the necessity of linearly separable data.

Multi-layer perceptron (MLP) network has shown itself as a powerful tool but some algorithms have big training times and high computer requirements. In certain cases, such as the Levenberg–Marquardt algorithm, these constraints make its use not viable. Among the tested algorithms, resistant and conjugate gradient algorithms were selected due to their great results, robustness and short time of training. Learning vector quantization network (LVQ) [7] and the recurrent Elman network [8] did not present such good results as the previous ones, most of all in hard experiments, being also not very quick in training steps.

Self organizing maps (SOM) [7] show very good clustering results, grouping customers with similar activity or behaviour. With SOM it is also possible to easily evaluate the results through the graphical representation of the maps. Different customer labels can be grouped by visual inspection. Applying the k-means and Davies – Bouldin index functions is possible to obtain the best clusters of the map. Unsupervised learning is another advantage of SOM, allowing to group labels automatically.

Radial basis function neural networks [9] have a good performance in the subset of hybrid neural networks (HNN) due to their lack of training steps, which makes them faster. Two methodologies stand out against HNN due to the great results achieved: the generalized regression neural network (GRNN) and probabilistic neural network (PNN) [10]. Specifically, the application and use of the last one is very easy, and this fact improves the evaluation of results. However, another radial base network, the RBF, shows some problems when the experiment grows in complexity.

Adaptive neuro-fuzzy inference system (ANFIS) [11] shows good results too, but the computing requirements, and so the training time, is too high when the number of data is considerable and thus the method becomes not recommendable.

Two fuzzy-logic techniques have been analysed: C-means and subtractive clustering. The first one for seg-

mentation purposes and the second one for classification and identification (generally for new input data); both have achieved great results. Indeed, the subtractive clustering, although is not a fast technique, has a remarkable robustness and quality.

In conclusion and after some preliminary tests, two different groups were found in this set of methodologies:

- Methodologies showing a considerable ability to classify and group the input space database, such as Fuzzy c-means clustering.
- Methodologies showing an ability to classify the input space database and furthermore to identify new customer patterns when new customers or measurement increase the database, i.e., memory behaviour. For example, MLP, RBF, GRNN, SOM, PNN, ANFIS, and subtractive clustering.

Inside this last group, it cannot be concluded that a certain approach is clearly better or worse than others. Each one has its pros and cons, but it is also true that some methodologies show great performances for the research interest: fast processing capacity, the quality of results when the problem attains a high level of complexity and the ability to learn from a database to produce a further classification and identification when the input space grows. SOM, GRNN and PNN show these attributes, but in this research work a preference for SOM tools is reported due to their better understanding capacity – when the operator is not an expert in ANN techniques- and, of course, the unsupervised training feature.

4. Self-Organizing Maps

The management of databases and the process that aims at extracting synthesized information from large amounts of data can be performed by a lot of techniques, such as relational statistical calculus – a traditional approach – or, for example, automatic learning – fuzzy inference or artificial neural networks. This last approach has been selected for this application, and specifically self-organizing maps (SOMs). The methodology was introduced by Teuvo 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 an output space – two dimensional in this case. The advantage of SOM is that the relation between the original vectors is to some extent preserved in the two dimensional space, providing – through some analysis of these maps and the evaluation of some indices – a visual format that allows a human operator, with some expertise, to "easily" discover clusters, relations and structures in these databases.

5. Data Processing and Conditioning

5.1 Generalities

The mining of information from a database needs a previous treatment of the data to give said data a uniform format that allows its handling. The daily load profiles present peaks and different values for each user. Therefore, it is necessary to standardize the data to have a common format for all users and to allow their introduction in the neural network or the corresponding method of study avoiding the levels of demand in the first approximation.

Previous studies allowed the authors to achieve the conclusion that the most suitable format for data was fitting with load curves that presented 24 daily values of consumption [12], as it is necessary to extract the more relevant characteristics of the user – with the minor cost of time and work – while considering the availability of the sources – measures usually taken by electrical companies. While for the filtering purpose data is normalized by means of monthly peak values so the anomalous days can be found easily, for the classification of customers data daily peak values are chosen so it can be aggregated to customers with the same load curve shape but with a different level of load in the same cluster.

5.2 Case Study

The daily load profiles that compose the set of measurements for training and evaluation of SOM maps correspond to a mix of industrial, institutional and small residential loads – in this last case the load is aggregated in the high voltage side of a distribution transformer centre, CT. The annual peak load varies from 200 kilowatts – for the smaller user – to 10MW as the greater customer demand. Table 1 shows the individual customers selected from different geographical areas to perform training (U1, U2, ...), the mask associated to daily load curves – a number of masks for all the load curves for a given customer, the number of load curves considered for each customer, and the first seen customer classification according to its activity – institutional residential and industrial. So the global number of n-dimensional vectors used for training SOM was 327, each one corresponding to a demand curve.

For the filtering explanation only two typical medium users, an industry and a university, were considered. Data used in the training of the neural network corresponds with weekly load curves – from Monday to Friday – of both customers measured in 2003. For simplicity, Saturdays and Sundays profiles were not considered in this phase of study as only anomalous days wanted to be appointed.

6. Application of SOM For Anomalous Data Filtering

To analyse the possibilities of SOM for load data filtering, two users were selected from the customer spectra: specifically, a university and a medium industry to which the greater number of load curves records belonged – several months – were available. Obviously, these records include some anomalous days and wrong measurements. Moreover, this number determines in a certain way the size of the network (see Table 2).

An alternative labelling is used for data filtering purposes. This filtering is applied separately to each customer. By means of this labelling 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 (i.e., mm/dd). Thus, the

Table 1
Customer Spectrum for SOM Training

Customer	Label	No. of Input n-Data Vectors	Customer Activity
U1	1	22	University
U2	2	16	University
I1	3	15	Industry
I2	4	15	Industry
R1	5	6	Residential CT
R2	6	5	Residential CT
R3	7	5	Residential CT
R4	8	5	Residential CT
U3	9	8	University
U4	10	7	University
U5	11	20	University
U6	12	20	University
U7	13	20	University
U8	14	20	University
U9	15	20	University
U10	16	22	University
I3	17	30	Industry
I4	18	22	Industry
R6	19	24	Residential CT
R7	20	25	Residential CT

Table 2
Customers Spectrum for SOM Training

Customer	No. of Initial Input Data Vectors	Customer Activity
User 1	147	Medium Industry
User 2	195	University

label map that is obtained (see Fig. 2(a)) allows the identification of daily load data assigned to each cell.

The information contained in the daily load curves can be represented by means of time domain or frequency domain parameters. In this case, time domain representation seems quite useful. Specifically, load curves recorded every 15 min were interpolated to obtain the 24-values daily consumption curves finally used. The reason was the good results obtained in previous works accomplished by the Valero *et al.* [12].

The simulations were made using a Pentium IV computer with 512 Megas Ram and with the Matlab software. Different possible configurations were proved for the parameters of the SOM network. The training times of SOM maps were just a few minutes. A map size of 16×16 cells and a number of 2000 and 1000 steps for primary and secondary training respectively was finally applied using the label map obtained by means of the criterion already explained (mm/dd). The anomalous days were clearly located (see again Fig. 2(a)).

For example, Fig. 2 shows how labels 501 (1st May) and 1,208 (8th December), corresponding to two holidays in Spain, were both located in the left bottom area of the map. Also a county holiday marked with label 1009 (9th September) is located closed to the previous ones. Obviously, the cells next to the previous one – 1,231, 1,225 – are festive days in Christmas Season.

Finally and once the network is trained, it is possible to force it to group data fixing an upper limit of clusters. When a maximum number of 10 clusters is allowed, the four zones defined in Fig. 2(b) are found.

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 profiles: typical consumption patterns, assigned to regions 3 and 4 (see Fig. 3); profiles placed in region 1, which due to the characteristic of topographic preservation of SOM are identified as holidays (i.e., anomalous days, see Fig. 4); and finally profiles that denote standards of different behaviour from the usual ones lie in region 2 (see Fig. 5).

The previous results show that the network is not only able to identify anomalous days with the consistent capacity of cleanliness of the database, but it also presents other usefulness or applications. Thus, the network isolates consumption profiles with erroneous measures caused by failure in demand meters and will enable the utilities to manage them correctly – to reconstruct them or to eliminate wrong data. Moreover, it allows the study of particular behaviours of the customer. So demand patterns of the studied university identified in cluster 2 correspond to days of July during which there are relative night-time demand level growths due to the continuous use of air conditioning and a day-time peak demand reduction owed to student's holidays period. Therefore, the exposed method is presented as a useful tool in the study and estimation of end use loads – air conditioning in this case – and its possible routes of management – control actions, introduction of new technologies, and efficiency measures.

7. Customer Patterns Classification

The first purpose of application of SOM is to obtain a map for the identification and classification of electrical customers with different demand patterns that may be later employed in diverse tasks [13–15]. First, an example of a simple classification task is presented using data of only two customers, showing the difference in training with and without the filtering step previously explained. Later in a second part, a training with a greater number of filtered customers will be presented.

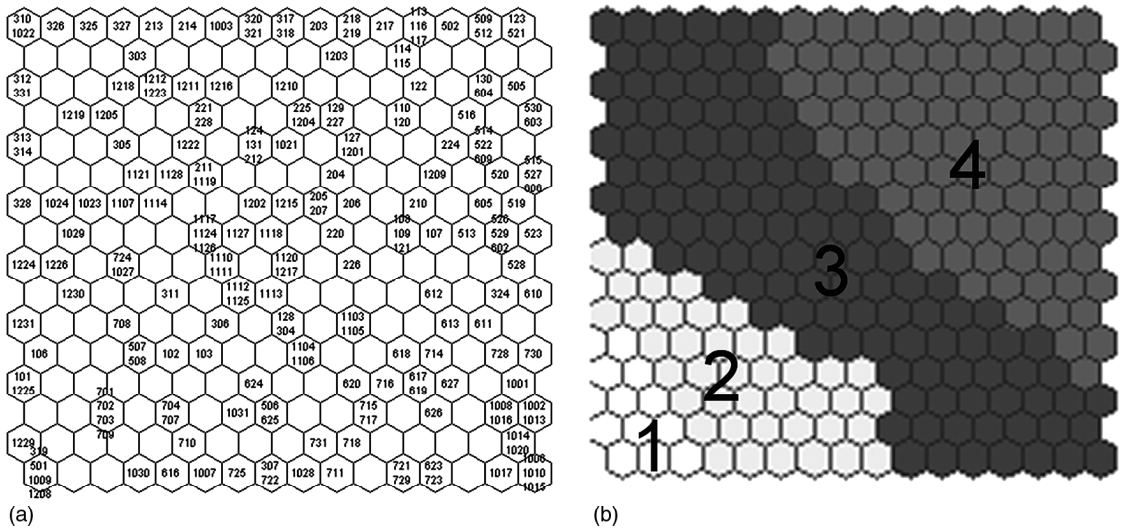


Figure 2(a) and (b). University label map (mm/dd criterion) and University clusters map with 10 maximum clusters.

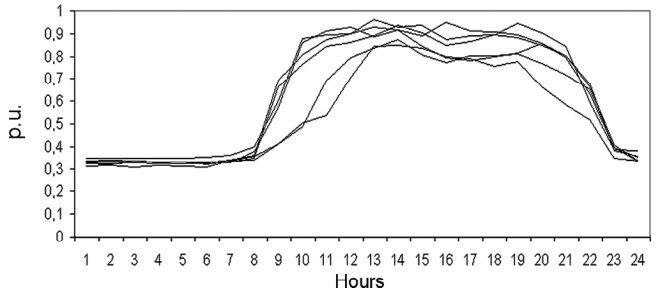


Figure 3. Typical consumption pattern. University (map regions 3 and 4).

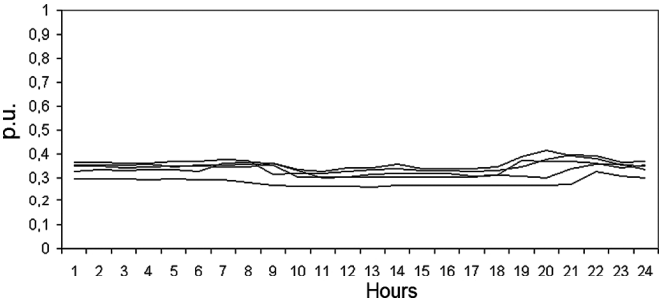


Figure 4. Holidays consumption pattern. University (map region 1).

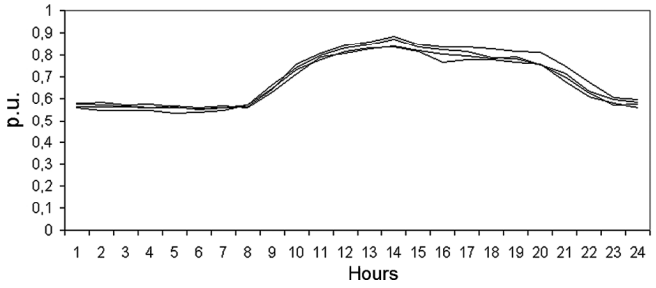


Figure 5. Change in patterns due to air conditioning. University (map region 2).

7.1 Preliminary Classification of Customer: The Advantages of Data Filtering for SOM Training

Subsequently, the data conditioning process previously described using the network's characteristics is shown in Section 5. A label map obtained using mask 1 for medium industry and mask 2 for the University can be seen in Fig. 6a. In this figure, it is shown how the network is able to group the customer's profiles in the same map area, separating and differentiating each one.

Nevertheless, it is shown in the same figure how the map presents zones where customer separation is not clear and some labels of both customers are even mixed. This fact suggests that some profiles belonging to different customers are not well suited to typical consumption patterns and therefore the network is not able to distinguish them adequately. For that reason, a new training was performed where the anomalous demand profiles are removed. In this case, the filtering was separately applied to both customers. The results can be seen in Fig. 6b: the network is now able to classify correctly each customer, and besides, it separates with higher resolution the data corresponding to both customers.

Typical customers' profiles classified by the SOM can be seen in Fig. 7. Notice the necessity of previous database filtering due to the similarities found among industry load profiles and university holiday profiles – a load curve with a higher load factor, (see Fig. 6a and b).

The improvement is also stated by the use of quantization (Qe) and topographic (Te) errors. The first error gives information of the accuracy with which the map represents the samples of data. Topographic error (Te) measures the percentage of profiles that have non-closed first and second winning units. As it can be seen, the value for the second one is negligible – a small value indicates that the network does not "hesitate" in the profiles assignation to cells. (see Table 3).

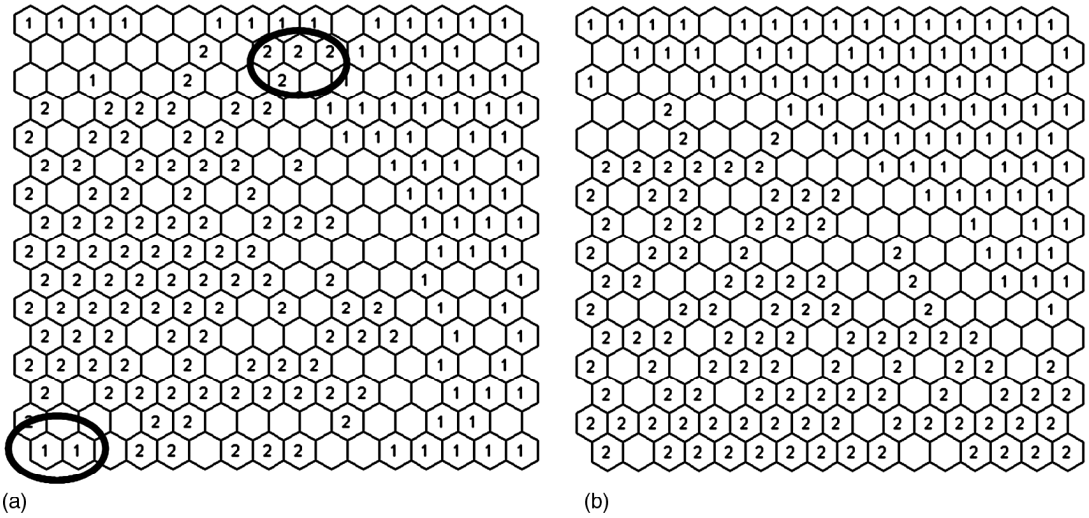


Figure 6(a) and (b). Customer classification without filtering and with previous filtering.

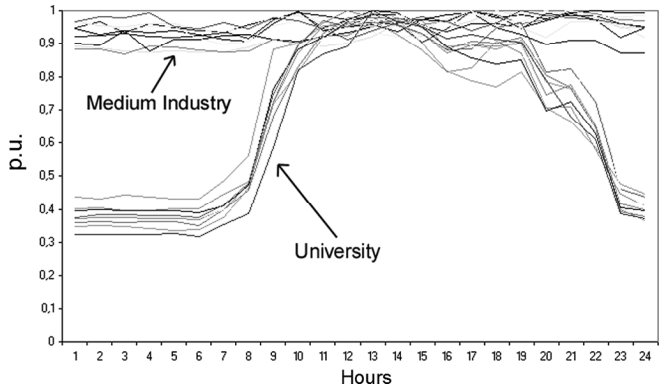


Figure 7. University and industry load profiles.

Table 3
Quantization and Topographic Errors

SOM Training	Qe	Te	Qe Relative Reduction (%)
With anomalous days	0.1520	0.0114	–
Without anomalous days	0.1303	0.0150	14.3

7.2 Overall Customer Sample Classification

7.2.1 Customer Aggregation

It has been described that load data filtering improves the results of the application of SOM to the classification of a reduced number of customers. The purpose of this paragraph is to test the capacity of SOM network to deal with a large number of customer load curves and to classify them in coherent segments. To do so, the filtered load curves of customers described in Table 1 were trained in a 21×21 cell map, with a primary and secondary training of 8,000 and 4,000 steps. The result is shown in Fig. 8.

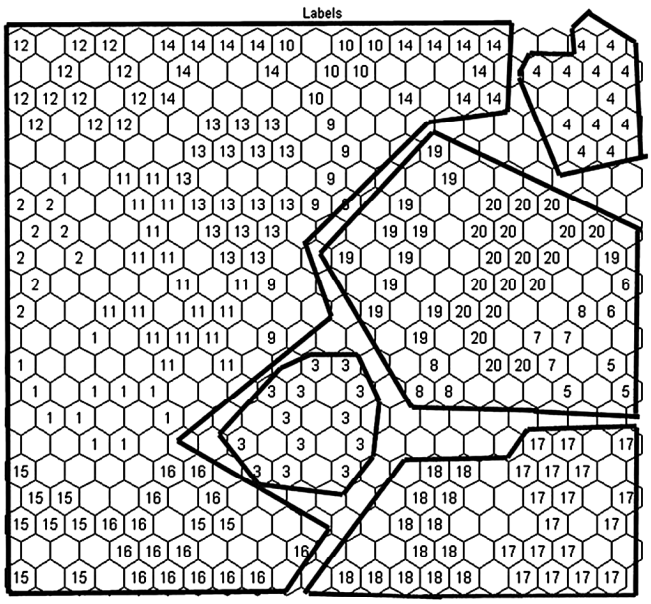


Figure 8. Customer aggregates.

Different zones of customers with similar characteristics – economical activity – can be seen in the map. The 20 customers simulated can be classified in five different zones: university (labels 1, 2, 9, 10, 11, 12, 13, 14, 15 and 16), residential (labels 5, 6, 7, 8, 19 and 20), medium industry (labels 3 and 4) and small industry (labels 17 and 18).

7.2.2 Index Matching Performance

As it has been stated before, the second objective of the research is to show the capacity of SOM for customer segmentation. Three new customers unknown by our SOM network – a university, a small industry, and a residential CT – (see Table 4) were used to test the SOM adequacy for customer classification. The target is to get for each new n-data input – load curve – the most similar cell (i.e., the segment of the new customer presented to the SOM map).

Table 4
Costumer Testing

Customer	Number of Entry Data Available (Number of Daily Load Curves)	Type of Customer
University A	16	University
Industry B	21	Small industry
Residential C	31	Residential CT

The results are presented in Table 5. The n-data of new customers were perfectly assigned by SOM to the customer segment they represent without failure, so SOM capability for customer identification and segmentation is stated.

Table 5
Results of Customers Testing

Customer Tested	Labels Identified	Kind of Customer Identified (%)	Index of Success
University A	9 cells with label 13	U7 56,25	1
	7 cells with label 15	U9 43,75	
Small industry B	21 cells with label 17	I3 100	1
Residential C	19 cells with label 5	R1 61,29	
	5 cells with label 20	R6 16,13	
	4 cells with label 19	R5 12,90	1
	3 cells with label 8	R4 9,68	

8. Conclusions

A SOM development is applied to distribution power system operation, and management is presented in this work to achieve the segmentation and demand patterns classification for electrical customers on the basis of their daily load curves. In the case where possible presence of anomalous data, and so uncertainty, appears – such as the case of large utility databases – this neural network tool also provides the effective detection of outliers, missing information or, e.g., excursions from standard pattern – due to price changes, or external factors as is the case of external temperature growth.

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, e.g., the potential of energy efficient alternatives, customer response to real price or TOU tariffs, tariff diversification, the success of dual-fuel or energy storage appliances or the possibilities of distributed generation in medium and small users. The research activity already in study is devoted to the development of two objectives: the improvement of segmentation indices used in the SOM map, including some external parameter, such as economical activity, quality and reliability of supply, tariffs, etc., and the development of new tools based in ANN to identify the percentage of a kind of customer in a distribution system. The results of these works will be reported by the authors in the near future.

Acknowledgment

This work was supported by Ministerio de Educación y Ciencia (Spanish Government) through research project ENE2007-67771-C02-02

References

[1] California Energy Commission, *Integrated Energy Policy Report,* Docket No 02-IEP-1, Pub No 100-03-019, December 2003.

[2] C. Olaru & L. Wehenkel, Data mining tutorial, *IEEE Computer Applications in Power*, *12*(3), 1999, 19–25.

[3] B.D. Pitt & D.S. Kirschen, Application of data mining techniques to load profiling, *Proc. IEEE PICA'99*, Santa Clara, CA, May 16–21, 1999, 131–136.

[4] G. Chicco, R. Napoli, F. Piglione, P. Postolache, *et al.*, Load pattern-based classification of electricity customers, *IEEE Transactions on Power Systems*, *19*(2), 2004, 1232–1238.

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

[6] F. Rosenblatt, *Principles of neurodynamics* (Washington DC: Spartan Press, 1961).

[7] T. Kohonen, *Self-organisation and associative memory*, Third Edition (Berlin: Springer-Verlag, 1989).

[8] J.L. Elman, Finding structure in time, *Cognitive Science*, *14*, 1990, 179–211.

[9] S. Chen, C.F.N. Cowan, & P.M. Grant, Orthogonal least squares learning algorithm for radial basis function networks, *IEEE Transactions on Neural Networks*, *2*(2), 1991, 302–309.

[10] D. Gerbec, S. Gasperic, I. Smon, & F. Gubina, Allocation of the load profiles to consumers using probabilistic neural networks, *IEEE Transactions on Power Systems*, *20*, 2005, 548–555.

[11] J.-S.R. Jang, ANFIS: Adaptive-network-based fuzzy inference system, *IEEE Transactions on Systems, Man and Cybernetics*, *23*, 1993, 665–685.

[12] S. Valero, M. Ortiz, Fco. García, A. Gabaldón, *et al.*, Characterization and identification of electrical customer through the use of SOM and daily load parameters, *IEEE PSCE2004*, New York, October 10–13, 2004.

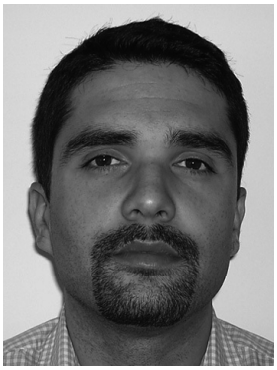
[13] G. Chicco, R. Napoli, & F. Piglione, Load pattern clustering for short-term load forecasting of anomalous days, *2001 IEEE Porto Power Tech Conf.*, 10–13 September, Porto, Portugal.

[14] C.S. Chen, J.C. Hwang, & Y.M. Zzeng, C.W. Huang, & M.Y. Cho, Determination of customer load characteristics by load survey system at taipower, *IEEE Transactions on Power Delivery*, *11*(3), July 1996, 1430–1436.

[15] R. Lamedica, A. Prudenzi, M. Sforna, M. Caciotta, *et al.*, A neural network based technique for short-term load forecasting of anomalous load periods, *IEEE Transactions on Power Systems*, *11*(4) 1996, 1749–1756.

Biographies

Sergio Valero Verdú was born in Elche, Spain, in 1974. He received his degree in Industrial Engineering in 1998, from the Universidad Politécnica de Valencia, Spain. Currently he is an Associate Professor at the Universidad Miguel Hernández de 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 his degree in Industrial Engineering in 2002, from the Universidad Politécnica de Cartagena, Spain.

Currently he is an Associate Professor at the Universidad Miguel Hernández de 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 Blanes received her degree in Engineering in 1998 from the Universidad Politécnica de Valencia. At the moment she is working at the Ph.D. degree in Industrial Engineering, at the Universidad Miguel Hernández de Elche. In 2001 she became an Associate Professor in Mechanical Engineering at the Universidad Miguel Hernández de Elche and has collaborated in several projects within the Electrical Engineering Area regarding the electricity markets. She has authored numerous publications and contributions to congresses.

Francisco J. García Franco was born in Cartagena, Spain, in 1979. He received his Industrial Engineering degree in Electrical Power Systems in 2003 from the Universidad Politécnica de Cartagena, Spain. Currently he is a research and Ph.D. student in the Institute of Energy Engineering at the Universidad Politécnica de Valencia. His research activities include Electricity Markets, Demand Modelling, Demand-Side Bidding and Electrical Customer Classification.

Antonio Gabaldón Marín was born in Cieza, Spain, in 1964. He received his Industrial Engineering degree in 1988, and Ph.D. from the Universidad Politécnica de Valencia, Spain, in 1991.

Currently he is a Full Professor at the Universidad Politécnica de Cartagena, Spain. His research activities include Distribution System Analysis, Electricity Markets, Demand Modelling, Distributed Energy Resources, Demand-Side Bidding and Demand-Responsiveness.