Two-Stage Pattern Recognition of Load Curves for Classification of Electricity Customers

George J. Tsekouras, *Member, IEEE*, Nikos D. Hatziargyriou, *Senior Member, IEEE*, and Evangelos N. Dialynas, *Senior Member, IEEE*

***Abstract*—This paper describes a two-stage methodology that was developed for the classification of electricity customers. It is based on pattern recognition methods, such as k-means, Kohonen adaptive vector quantization, fuzzy k-means, and hierarchical clustering, which are theoretically described and properly adapted. In the first stage, typical chronological load curves of various customers are estimated using pattern recognition methods, and their results are compared using six adequacy measures. In the second stage, classification of customers is performed by the same methods and measures, together with the representative load patterns of customers being obtained from the first stage. The results of the first stage can be used for load forecasting of customers and determination of tariffs. The results of the second stage provide valuable information for electricity suppliers in competitive energy markets. The developed methodology is applied on a set of medium voltage customers of the Greek power system, and the obtained results are presented and discussed.**

***Index Terms*—Adaptive vector quantization, chronological load patterns, clustering, customer classes, fuzzy k-means, hierarchical clustering, k-means, pattern recognition.**

I. Introduction

In a competitive electrical energy market, it is highly desirable for suppliers to know the electrical behavior of their customers, in order to provide them with satisfactory services at the least cost. Suppliers are usually allowed to change the applied tariffs under the supervision of the regulatory energy authorities. For this purpose, the suppliers cluster customers in representative groups and use class average load curves to study the customers' behavior and apply new tariffs. Clustering techniques can be also used for load forecasting.

The classification of customers is mainly based on activity-type parameters and *a priori* typical indices, such as maximum monthly peak load, load factor, etc. During the last years, significant research effort has been devoted to load pattern-based clustering for the adaptation of multiple rate types [1]. The electricity behavior can be expressed either by daily and weekly indicators, such as the fill-up coefficient (average load to maximum daily load [1]) or by mean load curves of specific time periods (workdays of a three-week period [2]). The clustering methods being used are the "modified follow the leader" [1]–[4], the self-organizing map [2], [4], [5], the k-means [4], [5], the average and Ward hierarchical methods [3], [4], [6], and the fuzzy k-means [3], [4], [6]–[8]. These methods generally belong to pattern recognition techniques [4]–[9]. In order to reduce the size of the clustering input data set, Sammon map, principal component analysis and curvilinear component analysis have been proposed [4]. The respective adequacy measures that are commonly used are the mean index adequacy [1]–[3], the clustering dispersion indicator [1]–[4], the similarity matrix indicator [3], the Davies–Bouldin indicator [3], [4], [7], [8], the modified Dunn index [4], the scatter index [4], and the mean square error [6]–[8]. Alternatively, the classification of customers can be performed by data mining [10], [11], wavelet packet transformation [12], and load survey with stratified sampling [13].

The objective of this paper is to present a new two-stage methodology for customer's classification, which defines both the typical load curves of each customer and the customer clusters together with their respective representative load curves. In the first stage, the load curves of each customer are organized into well-defined and separated classes, in order to successfully describe the respective electricity customer's behavior. This allows the selection of adequate tariffs and the successful application of demand-side management programs. In the second stage, the classification of customers is carried out by using the characteristic typical load curve of each customer, being obtained from the first stage. In both stages, the method compares the results obtained by certain clustering techniques [k-means with special weights initialization, adaptive vector quantization (AVQ), fuzzy k-means, and seven hierarchical agglomerative clustering methods] using six adequacy measures (mean square error, mean index adequacy, clustering dispersion indicator, similarity matrix, Davies–Bouldin indicator, ratio of within cluster sum of squares to between cluster variation). Results from the application of the developed methodology on a set of medium voltage customers of the Greek distribution system are presented.

II. Two-Stage Pattern Recognition Methodology for Classification of Customers

The classification of customers is achieved by applying a pattern recognition methodology consisting of two stages, which has the following basic steps, and its flow chart is shown in Fig. 1.

Manuscript received August 8, 2006; revised October 27, 2006. This work was supported in part by the General Secreteriat for Research and Technology of Greece, European Union, and Greek Public Power Cooperation under Grant 97 YP 8. Paper no. TPWRS-00512-2006.

The authors are with the School of Electrical and Computer Engineering, National Technical University of Athens, Athens, Greece (e-mail: tsekouras_george_j@yahoo.gr; nh@power.ece.ntua.gr; dialynas@power.ece.ntua.gr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TPWRS.2007.901287

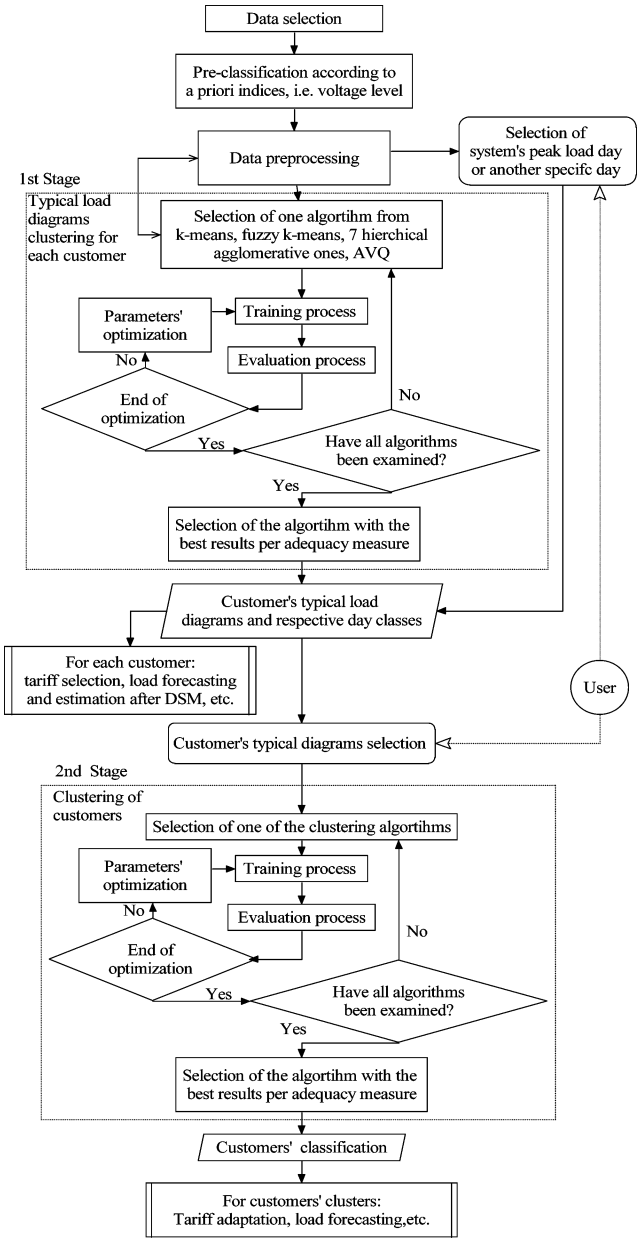

Fig. 1. Flow chart of two-stage pattern recognition methodology for the classification of customers.

1) *Data and features selection*: Using electronic meters for medium and high voltage customers, the active and reactive energy values are registered (in kWh and kVARh) for each time period in steps of 15 min, 1 h, etc. The daily chronological load curves for each individual customer are determined for each study period (month, season, year).
2) *Customers' clustering using a priori indices*: Customers can be characterized by their geographical region, voltage level (high, medium, low), economic activity, installed power, contracted energy, power factor, etc. These indices are not necessarily related to the load curves according to the experience of the Greek power distribution company. They can be used however for the pre-classification of customers. It is mentioned that the load curves of each customer are normalized using the respective minimum and maximum loads of the period under study.
3) *Data preprocessing*: The load curves of each customer are examined for normality, in order to modify or delete the values that are obviously wrong (*noise suppression*). If necessary, a preliminary execution of a pattern recognition algorithm is carried out, in order to track bad measurements or networks faults, which, if uncorrected, will reduce the number of the useful typical days for a constant number of clusters.
4) *Typical load curves clustering for each customer—First stage application of pattern recognition methods*: For each customer, a number of clustering algorithms (k-means, adaptive vector quantization, fuzzy k-means, and hierarchical clustering) is applied. Each algorithm is trained for the set of load curves and evaluated according to six adequacy measures. The parameters of the algorithms are optimized, if necessary. The developed methodology uses the clustering methods that provide the most satisfactory results. This process is repeated for the total set of customers under study. Special customers, such as seasonal ones (e.g., oil-press industry, small seaside hotels) are identified.
5) *Selection of typical chronological load curves for customers*: The typical load curves of customers that will be used for the final clustering are selected by choosing the type of typical day (such as the most populated day, the day with the peak demand load or with the maximum demand energy, etc.). It is possible to omit the customer's typical load curves clustering, if the user wishes to compare the customer's behavior in specific days, such as the day of system peak load, the mean July workday, etc. However, the customers' behavior is not entirely representative for the period under study. It is noticed that special customers can be handled separately.
6) *Clustering of customers—Second stage application of pattern recognition methods*: The clustering methods are applied for the set of the customer's characteristic load curves. After algorithms' calibration, the clusters of customers and the respective classes representative load curves are formed.

III. Mathematical Modeling of Clustering Methods and Adequacy Measures

The mathematical modeling is the same either for chronological typical load curves of a simple customer (when N analytical daily load curves are given) or for clustering of N customers. The solution in both of these cases is given through clustering techniques, as the main objective is to obtain sets of days or customers, respectively, and load patterns.

For this purpose, N is defined as the population of the input vectors to be clustered. The $\vec{x}_\ell = (x_{\ell 1}, x_{\ell 2}, \ldots x_{\ell i}, \ldots x_{\ell d})^T$ symbolizes the ℓth input vector and d is its dimension. The corresponding set is given by $X = \{\vec{x}_\ell : \ell = 1, \ldots, N\}$. The values of $x_{\ell i}$ are normalized in the range [0, 1] using the upper $x_{\max}$ and lower $x_{\min}$ values of all elements of the original input patterns set according to

$$\hat{x}_{\ell i} = (x_{\ell i} - x_{\min})/(x_{\max} - x_{\min}). \quad (1)$$

Each classification process partitions the initial N input vectors to M clusters, which can be the typical days of the customer under study (first stage of the developed methodology) or the customer classes (second stage). The jth cluster has a representative, which is the respective load profile and is represented by the vector $\vec{w}_j = (w_{j1}, w_{j2}, \ldots, w_{ji}, \ldots w_{jd})^T$ of d dimension. This vector expresses the cluster's center or the weight vector, if a clustering artificial neural network is used. The corresponding set is the classes' set, which is defined by $W = \{\vec{w}_j, j = 1, \ldots M\}$. The subset of input vectors $\vec{x}_\ell$, which belong to the jth cluster, is Ω_j and the respective population of load curves is N_j. For the assessment of the classification algorithms, the following forms of distances are defined:

1) The distance between the representative vector $\vec{w}_j$ of jth cluster and the subset Ω_j that is calculated as the geometric mean of the Euclidean distances $d(\vec{w}_j, \vec{x}_\ell)$ between $\vec{w}_j$ and each member $\vec{x}_\ell$ of Ω_j:

$$d(\vec{w}_j, \Omega_j) = \sqrt{\sum_{\vec{x}_\ell \in \Omega_j} d^2(\vec{w}_j, \vec{x}_\ell) \Big/ N_j}. \tag{2}$$

2) The infra-set mean distance of a set that is defined as the geometric mean of the inter-distances between the members of the set, i.e., for the subset Ω_j:

$$\hat{d}(\Omega_j) = \sqrt{\frac{1}{2N_j} \sum_{\vec{x}_\ell \in \Omega_j} d^2(\vec{x}_\ell, \Omega_j)}. \tag{3}$$

The basic characteristics of the four clustering methods being used are the following.

A. K-Means

The k-means clustering method groups the set of the N input vectors to M clusters using an iterative procedure. Initially the weights of the M clusters are determined. In the classical model, a random choice among the input vectors is used [4], [5]. In the developed algorithm, the w_{ji} of the jth center is initialized as

$$w_{ji}^{(0)} = a + b \cdot (j-1)/(M-1) \tag{4}$$

where a and b are properly calibrated parameters.

During epoch t for each training vector $\vec{x}_\ell$, its Euclidean distances $d(\vec{x}_\ell, \vec{w}_j)$ are calculated for all centers. The ℓth input vector is put in the set $\Omega_j^{(t)}$, for which the distance between $\vec{x}_\ell$ and the respective center is minimum. When the entire training set is formed, the new weights of each center are calculated as

$$\vec{w}_j^{(t+1)} = \frac{1}{N_j^{(t)}} \sum_{\vec{x}_\ell \in \Omega_j^{(t)}} \vec{x}_\ell \tag{5}$$

where $N_j^{(t)}$ is the population of the respective set $\Omega_j^{(t)}$ during epoch t. This process is repeated until the maximum number of iterations is used or the weights do not significantly change. The algorithm's main purpose is to minimize the appropriate error function. The main difference with the classical model is that the process is repeated for different pairs of (a, b).

B. Adaptive Vector Quantization

This algorithm is a variation of the k-means method, which belongs to the unsupervised one-layer neural networks. It classifies input vectors into clusters by using a competitive layer with a constant number of neurons. During epoch t, each input vector $\vec{x}_\ell$ is randomly presented and its respective Euclidean distances from every neuron are calculated. The weights of the winning neuron (with the smallest distance) are updated as

$$\vec{w}_j^{(t)}(n+1) = \vec{w}_j^{(t)}(n) + \eta(t) \cdot \left(\vec{x}_\ell - \vec{w}_j^{(t)}(n)\right) \tag{6}$$

where n is the number of input vectors, which have been presented during the current epoch, $w_{ji}^{(0)} = 0.5, \forall j, i$, and $\eta(t)$ is the learning rate according to

$$\eta(t) = \eta_0 \cdot \exp(-t/T_{\eta 0}) > \eta_{\min} \tag{7}$$

where η_0, $\eta_{\min}$, and $T_{\eta 0}$ are the initial value, the minimum value, and the time parameter, respectively. The remaining neurons are unchangeable for $\vec{x}_\ell$, as introduced by the Kohonen winner-take-all learning rule [14], [15]. This process is repeated until the maximum number of epochs is reached, the weights converge, or the appropriate error function does not improve.

C. Fuzzy k-Means

Each input vector $\vec{x}_\ell$ does not belong to only one cluster, but it participates to every jth cluster by a membership factor $u_{\ell j}$, where

$$\sum_{j=1}^{M} u_{\ell j} = 1 \;\&\; 0 \le u_{\ell j} \le 1, \forall j. \tag{8}$$

Theoretically, the membership factor gives more flexibility in the vector's distribution. The membership factors and the cluster centers are calculated in each epoch as

$$u_{\ell j}^{(t+1)} = 1 \Big/ \sum_{k=1}^{M} \frac{d\left(\vec{x}_\ell, \vec{w}_j^{(t)}\right)}{d\left(\vec{x}_\ell, \vec{w}_k^{(t)}\right)} \tag{9}$$

$$\vec{w}_j^{(t+1)} = \left(\sum_{\ell=1}^{N} \left(u_{\ell j}^{(t+1)}\right)^q \cdot \vec{x}_\ell\right) \Big/ \sum_{\ell=1}^{N} \left(u_{\ell j}^{(t+1)}\right)^q \tag{10}$$

where q is the *amount of fuzziness* in the range $(1,\infty)$ that increases as fuzziness reduces. The weights of the clusters centers are initialized by (4), which is similar to the developed k-means.

D. Hierarchical Agglomerative Algorithms

Agglomerative algorithms are based on matrix theory [9]. The input is the $N \times N$ dissimilarity matrix P_0. At each level t, when two clusters are merged into one, the size of the dissimilarity matrix P_t becomes $(N-t) \times (N-t)$. Matrix P_t is obtained from P_{t-1} by deleting the two rows and columns that correspond to the merged clusters and adding a new row and a new column that contain the distances between the newly formed cluster and the old ones. The distance between the newly

formed cluster C_q (the result of merging C_i and C_j) and an old cluster C_s is determined as

$$d(C_q, C_s) = \begin{cases} a_i \cdot d(C_i, C_s) + a_j \cdot d(C_j, C_s) + b \cdot d(C_i, C_j) \\ +c \cdot |d(C_i, C_s) - d(C_j, C_s)| \end{cases} \tag{11}$$

where a_i,a_j, b, and c correspond to different choices of the dissimilarity measure. It is noted that in each level t, the respective representative vectors are calculated by equation (4). The seven algorithms being used in the developed methodology are the single link (SL), the complete link (CL), the unweighted pair group method average $(UPGMA)$, the weighted pair group method average $(WPGMA)$, the unweighted pair group method centroid $(UPGMC)$, the weighted pair group method centroid $(WPGMC)$, and the minimum variance or $WARD$ algorithm [9].

E. Adequacy Measures

In order to evaluate the performance of the clustering algorithms and to compare them with each other, six different adequacy measures are applied. Their purpose is to obtain well-separated and compact clusters that make the load curves well identified.

1) *Mean square error or error function* (J) [6]:

$$J = \frac{1}{N} \sum_{\ell=1}^{N} d^2 \left(\vec{x}_\ell, \vec{w}_{k:\vec{x}_\ell \in \Omega_k} \right). \tag{12}$$

2) *Mean index adequacy* (MIA) [1]. It is defined as the average of the distances between each input vector assigned to the cluster and its center

$$MIA = \sqrt{\frac{1}{M} \sum_{j=1}^{M} d^2(\vec{w}_j, \Omega_j)}. \tag{13}$$

3) *Clustering dispersion indicator* (CDI) [1]. It is the ratio of the mean infra-set distance between the input vectors in the same cluster and the infra-set distance between the class representative load curves:

$$CDI = \sqrt{\frac{1}{M} \sum_{k=1}^{M} \hat{d}^2(\Omega_k)} \Bigg/ \hat{d}(W). \tag{14}$$

4) *Similarity matrix indicator* (SMI) [3]. It is defined as the maximum off-diagonal element of the symmetrical similarity matrix, whose terms are calculated by using a logarithmic function of the Euclidean distance between any kind of class representative load curves:

$$SMI = \max_{p>q} \left\{ \left(1 - 1/\ln\left[d(\vec{w}_p, \vec{w}_q)\right]\right)^{-1} \right\} : p, q = 1, \ldots, M. \tag{15}$$

5) *Davies–Bouldin indicator* (DBI) [16]. It represents the system-wide average of the similarity measures of each cluster with its most similar cluster:

$$DBI = \frac{1}{M} \sum_{k=1}^{M} \max_{p \neq q} \left\{ \left(\hat{d}(\Omega_p) + \hat{d}(\Omega_q) \right) \Big/ d(\vec{w}_p, \vec{w}_q) \right\} : \quad p, q = 1, \ldots, M. \tag{16}$$

TABLE I
COMPARISON OF THE BEST CLUSTERING METHODS FOR TEN CLUSTERS FOR A MEDIUM VOLTAGE INDUSTRIAL CUSTOMER

Adequacy measure	J	MIA	CDI	SMI	DBI	WCBCR
Classical k-means	0.254	0.0644	0.342	0.731	2.417	0.00672
Developed k-means	0.253	*0.0583*	*0.324*	0.671	*1.652*	*0.00668*
a parameter	0.10	0.19	0.35	0.17	0.18	0.11
b parameter	0.77	0.35	0.55	0.48	0.37	0.60
AVQ	*0.250*	0.0647	0.354	0.716	2.188	0.00689
η_0 - $T_{\eta 0}$	0.80-500	0.85-1500	0.85-500	0.75-2000	0.60-500	0.70-500
Fuzzy (q=6)	0.358	0.0714	0.370	0.764	3.256	0.00715
a parameter	0.31	0.27	0.18	0.13	0.10	0.10
b parameter	0.49	0.31	0.36	0.54	0.59	0.46
WARD	0.254	0.0680	0.372	0.730	2.733	0.00840
UPGMA	0.313	0.0630	0.401	*0.649*	2.171	0.00668
UPGMC	0.415	0.0766	0.435	*0.649*	2.120	0.00991

6) *Ratio of within cluster sum of squares to between cluster variation* $(WCBCR)$ [17]. It is the ratio of the sums of the distance's square between each input vector and its cluster's representative vector and the distances of the clusters' centres:

$$WCBCR = \sum_{k=1}^{M} \sum_{\vec{x}_\ell \in \Omega_k} d^2(\vec{w}_k, \vec{x}_\ell) \Bigg/ \sum_{1 \leq q < p}^{M} d^2(\vec{w}_p, \vec{w}_q). \tag{17}$$

The success of the different algorithms for the same final number of clusters is expressed by having small values of the adequacy measures. By increasing the number of clusters all the measures decrease, except of the similarity matrix indicator. An additional adequacy measure could be the number of the *dead* clusters, which contain empty sets. It is intended to minimize this number. It is noted that in (12)–(17), M is the number of the clusters without the dead ones.

IV. APPLICATION OF THE FIRST STAGE OF THE METHODOLOGY TO A SET OF MEDIUM VOLTAGE CUSTOMERS

The developed methodology was applied on the data concerning 94 medium voltage customers of the Greek distribution system whose maximum peak load ranges between 250 kW and 10 MW. These data are the 15-min load values for each individual customer for a period of ten months in the year 2003.

In the first stage, the typical load curves of each customer are determined. The analysis of one industrial customer is presented in detail, while additional customers can be analyzed in a similar way. The respective set of the daily chronological curves has 301 members. Nine curves were rejected through data preprocessing, while the remaining 292 curves were used by the above-mentioned clustering methods. In Table I, the results of the best clustering methods are presented for ten clusters.

A. Application of the k-Means

The developed model of the k-means method was applied for different pairs (a, b) from 2 to 25 clusters, where $a = \{0.1, 0.11, \ldots, 0.45\}$ and $a + b = \{0.54, 0.55, \ldots, 0.9\}$. For each number of clusters, 1332 different pairs (a, b) are examined. The best results for the six adequacy measures do not refer to the same pair (a, b). From the results of Table I, it is

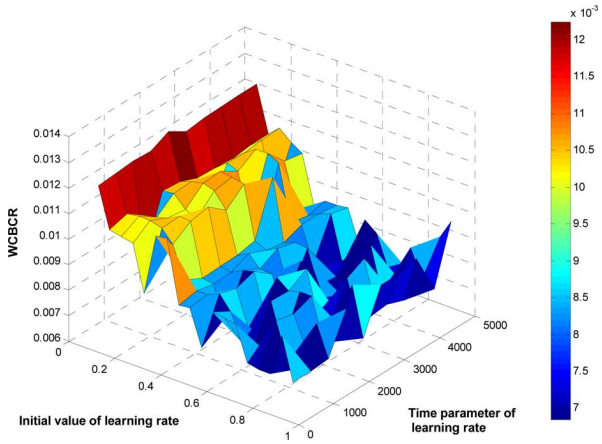

Fig. 2. WCBCR for the AVQ method for a set of 292 training patterns for ten neurons, $\eta_0 = \{0.1, 0.15, \ldots, 0.9\}$, $T_{\eta 0} = \{500, 1000, \ldots 5000\}$, $\eta_{\min} = 5 \cdot 10^{-6}$.

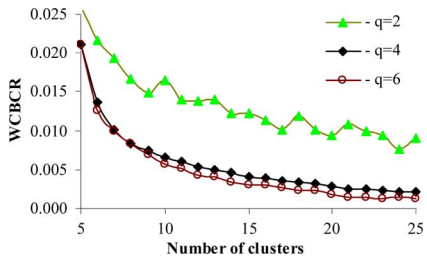

Fig. 3. WCBCR for the fuzzy k-means method for a set of 292 training patterns for 5 to 25 neurons for $q = 2, 4, 6$.

obvious that the developed k-means is superior to the classical one with the random choice of the input vectors during the centers' initialization [4], [5]. For the classical k-means model, 100 executions were carried out and the best results for each adequacy measure were registered. The superiority of the developed model applies in all cases of neurons. A second advantage is the convergence to the same results for the respective pairs (a, b), which cannot be achieved using the classical model.

B. Application of the Adaptive Vector Quantization

The initial value η_0, the minimum value $\eta_{\min}$, and the time parameter $T_{\eta 0}$ of learning rate must be properly calibrated. For example, in Fig. 2, the sensitivity of the ratio of within cluster sum of squares to between cluster variation WCBCR to the η_0 and $T_{\eta 0}$ parameters is presented for 90 experiments. The best results of the adequacy measures are not given for the same η_0 and $T_{\eta 0}$, according to the results of Table I. The $\eta_{\min}$ value does not practically improve the neural network's behavior, assuming that it ranges between 10^{-5} and 10^{-6}.

C. Application of the Fuzzy k-Means

In the fuzzy k-means algorithm, the results of all the adequacy measures improve, as the amount of fuzziness increases. As an example, the WCBCR is presented in Fig. 3 for different number of clusters for three cases of $q = \{2, 4, 6\}$.

D. Application of Hierarchical Agglomerative Algorithms

The best results are given by the WARD model for J and CDI adequacy measures and by the UPGMA model for MIA, SMI, WCBCR adequacy measures. For the Davies–Bouldin indicator, there are significant variances, as shown in Fig. 4.

E. Comparison of Clustering Models and Adequacy Measures

The best results for each clustering method (modified k-means, fuzzy k-means, adaptive vector quantization, and hierarchical algorithms) are presented in Fig. 5. The k-means model has the smallest values for the MIA, CDI, DBI, and WCBCR measures. The optimized AVQ presents the best behavior for the mean square error J and the unweighted pair group method average algorithm (UPGMA) for the SMI. The developed k-means model has similar behavior to the WARD algorithm for the J and to the UPGMA algorithm for the WCBCR. All measures (except DBI) show an improved performance, as the number of clusters increases.

The training time for the methods under study has the ratio 0.05:1:24:36 (hierarchical: proposed k-means: optimized adaptive vector quantization: fuzzy k-means for q = 6). Therefore, the k-means and hierarchical models were selected to be used. The number of dead clusters for the developed k-means model is shown in Fig. 6.

It is obvious that the use of WCBCR is slightly better than MIA and J measures. It is also noted that the basic theoretical advantage of the WCBCR measure is that it combines the distances of the input vectors from the representative clusters and the distances between clusters (covering also the J and CDI characteristics).

F. Representative Load Curves of a Customer

The number of clusters, which provides satisfactory results for the study of load demand behavior of a specific customer, corresponds to the knee of the curve [4], [7], [8]. If this knee is not clearly shown, the tangents are drawn, as shown in Fig. 7, estimating the knee for ten clusters. Alternatively, the best number of clusters can be determined, when further increment of clusters does not improve the respective adequacy measure significantly.

The results of the respective clustering for ten clusters using the developed k-means model with the optimization of the WCBCR adequacy measure are presented in Table II and in Fig. 8. The load demand behavior of the industrial customer during the year can be observed through the respective curves. Cluster 1 represents holidays, cluster 2 the first working day after holidays, clusters 9 and 10 the usual workdays, where every 8 h, there is a small variance due to the change of the shift.

During the training process of any clustering method, each measured clustered diagram is automatically assigned to the respective typical load curve of the customer, as it is shown in the example of Fig. 8. If the measured clustered diagram is new and it has not participated in the training process, the Euclidean distances from each typical load diagram of the customer are calculated, the smallest distance is found, and the measured clustered diagram is assigned to the respective typical load diagram. Additionally, in Fig. 8, the confidence limits of the variations

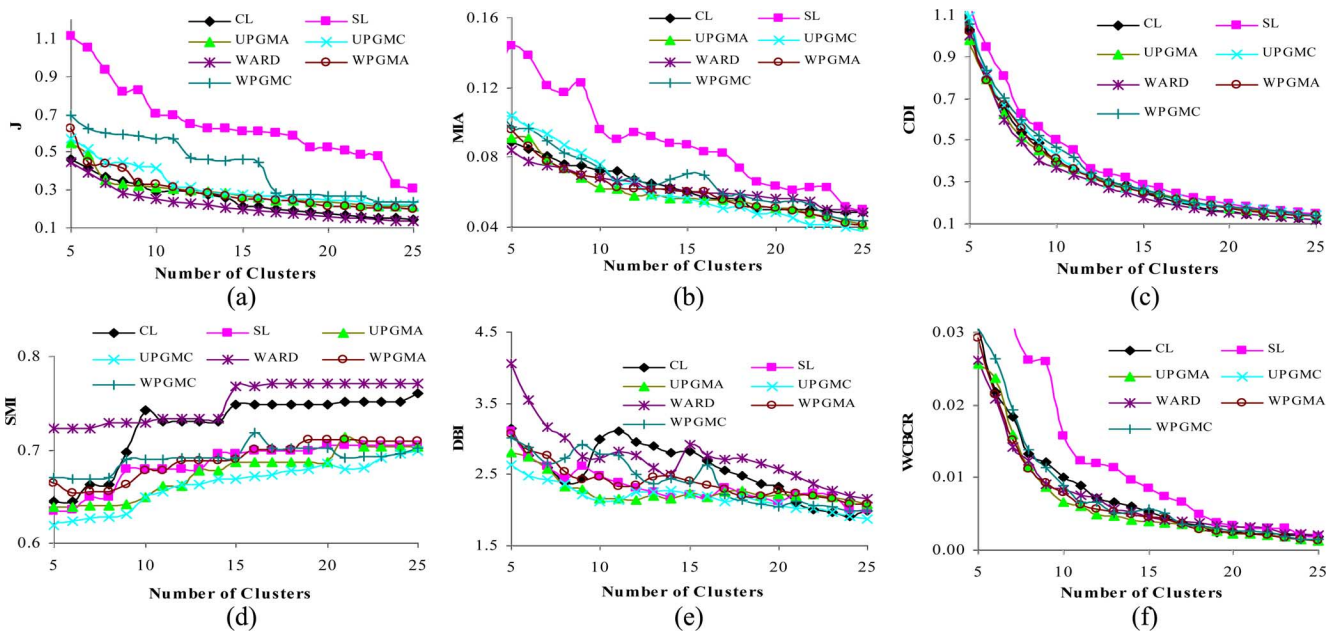

Fig. 4. Adequacy measures for the seven hierarchical algorithms for a set of 292 training patterns of a medium voltage industrial customer for 5 to 25 clusters. (a) J (similar to CDI). (b) MIA (similar to $WCBCR$). (c) CDI. (d) SMI. (e) DBI. (f) $WCBCR$.

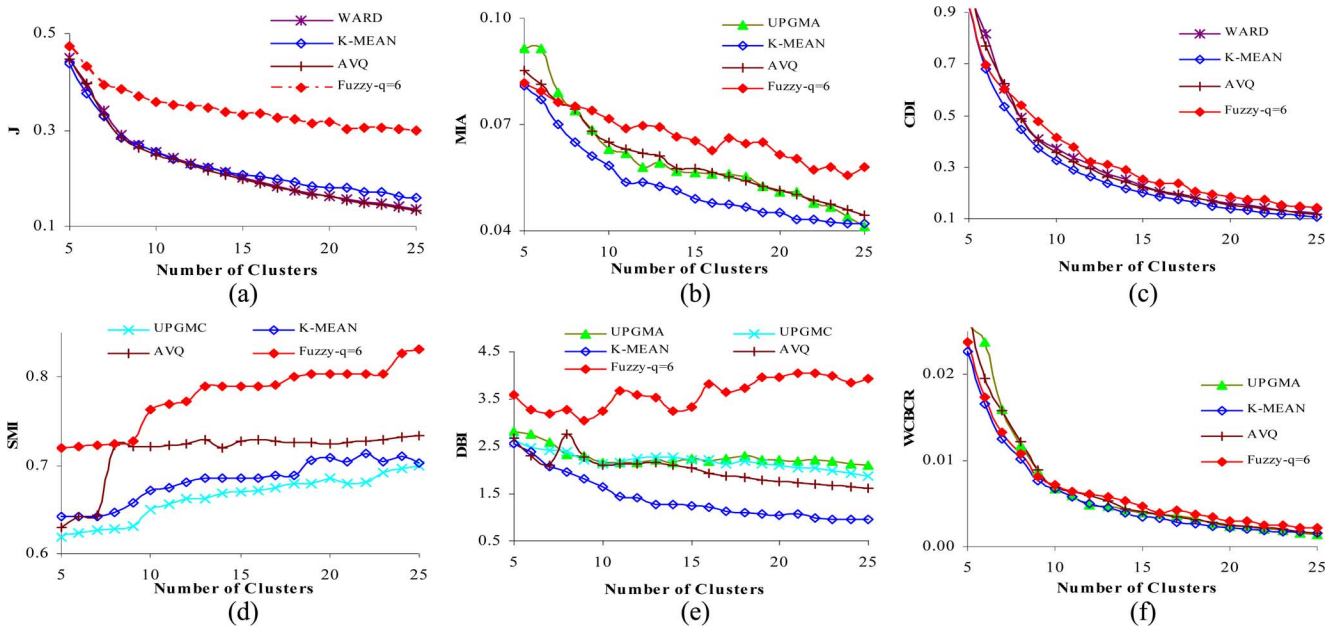

Fig. 5. The best results of each clustering method for the set of 292 training patterns of a medium voltage industrial customer for 5 to 25 clusters. (a) J. (b) MIA. (c) CDI. (d) SMI. (e) DBI. (f) $WCBCR$.

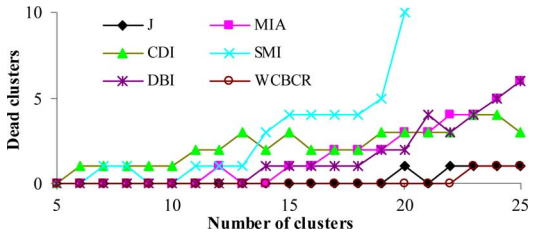

Fig. 6. Dead clusters for the developed k-means method optimizing the respective adequacy measure for the set of 292 training patterns of an MV industrial customer for 5 to 25 clusters.

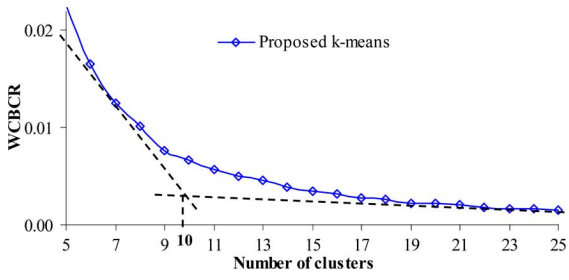

Fig. 7. WCBCR measure of the k-mean model for 5 to 25 clusters for the set of 292 training patterns of an MV industrial customer and the use of the tangents for the estimation of the knee.

(mean value $\pm$ standard deviation) are presented, and this has a probability of occurrence equal to 68.27% assuming a normal

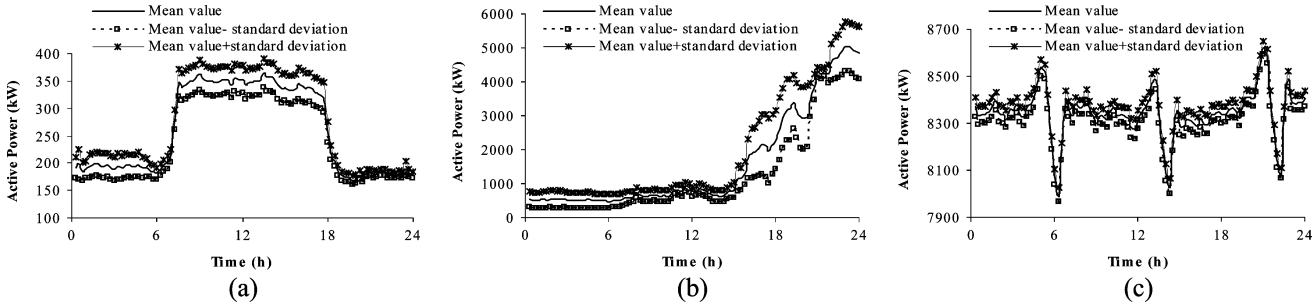

Fig. 8. Three examples of the characteristic daily chronological load curves for an MV industrial customer, where $\min \rightarrow 88.0\,\mathrm{kW}$, $\max \rightarrow 9468.8\,\mathrm{kW}$. (a) Cluster 1. (b) Cluster 2. (c) Cluster 10.

TABLE II
RESULTS OF THE DEVELOPED k-MEANS MODEL WITH OPTIMIZATION TO WCBCR MEASURE FOR TEN CLUSTERS FOR AN MV INDUSTRIAL CUSTOMER

Cluster	Day (1 for Monday, etc.)							Days per cluster
	1	2	3	4	5	6	7	
1	4	4	4	4	3	4	4	27
2	1	0	0	0	2	0	0	3
3	1	0	0	0	0	1	0	2
4	0	1	1	0	0	0	0	2
5	0	1	0	0	0	0	0	1
6	4	2	4	4	3	2	2	21
7	6	4	3	4	4	2	0	23
8	2	1	1	1	2	2	1	10
9	9	11	16	11	11	11	13	82
10	14	17	12	18	16	21	22	120
Total number of days under study								292

TABLE III
COMPARISON OF THE CLUSTERING MODELS FOR THE SET OF 94 MEDIUM VOLTAGE CUSTOMERS FOR TEN CLUSTERS

Adequacy measure	J	MIA	CDI	SMI	DBI	WCBCR
Classical k-means	0	0	0	0	0	0
Proposed k-means	26	28	94	24	18	37
AVQ	26	2	0	36	50	1
Fuzzy q=6	0	0	0	0	2	6
CL	0	1	0	0	0	0
SL	0	5	0	2	5	4
UPGMA	0	17	0	4	8	12
UPGMC	0	21	0	25	6	18
WARD	42	1	0	0	1	1
WPGMA	0	12	0	1	3	8
WPGMC	0	7	0	2	2	7

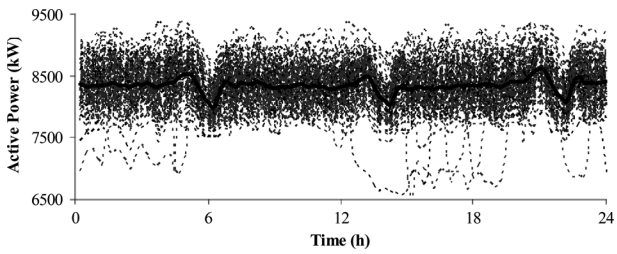

Fig. 9. Daily chronological load curve of cluster 10 for the MV industrial customer (bold line) along with its 120 clustered measured curves (thin lines).

distribution. In Fig. 9, the typical load curve of cluster 10 (which represents the most populated day) along with the 120 measured clustered load curves are shown. It is obvious that the number of the days for each representative cluster of this customer is not influenced by the day of the week.

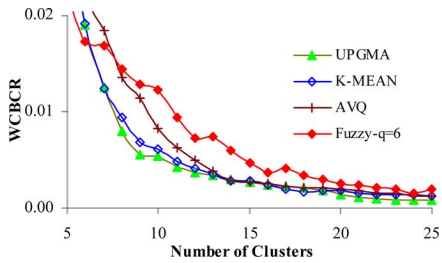

Fig. 10. WCBCR measure of the best fitting clustering methods for 5 to 25 neurons for the training patterns set of 94 medium voltage customers.

G. Application of Clustering Algorithms to All Customers

The same process was repeated for the remaining 93 customers where the load curves of each customer are qualitatively described by using 8–12 clusters. A different number of clusters for different customers can be estimated. The performance of these methods is presented in Table III by indicating the number of customers that achieve the best value of adequacy measure, respectively. The comparison of the algorithms showed that the developed k-means method achieves a better performance for MIA, CDI, and WCBCR measures, the optimized AVQ model for SMI and DBI measures, and the Ward model for J measure. It can be noticed that the classical k-means model shows the worst performance in adequacy measures.

V. APPLICATION OF THE SECOND STAGE OF THE METHOD TO A SET OF MEDIUM VOLTAGE CUSTOMERS

The customers' classification and the respective typical load curves are calculated. Additionally, the clustering methods are applied for the set of the characteristic customer's typical load curves for the 94 medium voltage customers. The characteristic customer's typical day can be either the most populated day of the customer or the day with the peak load demand (independently of the best number of clusters for each individual customer). For this case study, the most populated day of each customer is used (cluster 10 for the customer in Table II). The representative load curve for each customer is obtained by the clustering method that shows the best results for the adequacy measure being used.

Fig. 10 shows the best results of each clustering method by using the WCBCR measure. The developed k-means and

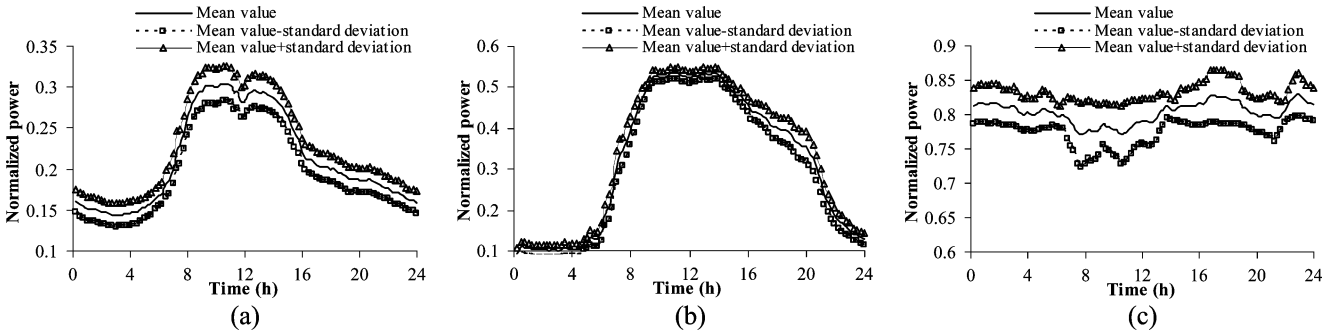

Fig. 11. Three indicative examples of the representative chronological load curves for a set of 94 medium voltage customers. (a) Cluster 3. (b) Cluster 5. (c) Cluster 12.

TABLE IV
RESULTS OF THE UPGMA MODEL WITH OPTIMIZATION TO WCBCR MEASURE FOR 12 CLUSTERS FOR A SET OF 94 CUSTOMERS

Load cluster	Activity of customer (1: industrial, 2: commercial, 3: public services, 4:traction)				Customers per cluster
	1	2	3	4	
1	4	5	0	0	9
2	0	1	0	0	1
3	8	16	3	2	29
4	1	0	0	0	1
5	1	13	3	1	18
6	2	7	0	0	9
7	3	9	0	0	12
8	2	2	1	0	5
9	1	0	0	0	1
10	1	0	0	0	1
11	3	0	0	0	3
12	5	0	0	0	5
Total	31	53	7	3	94

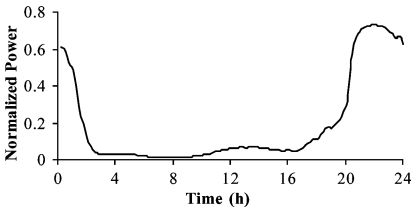

Fig. 12. Representative chronological load curve of the separate customer of cluster 2 for a set of 94 medium voltage customers.

UPGMA models were proved to be the best ones. The results of clustering for 12 clusters using the UPGMA model with the optimization of the WCBCR measure are presented in Table IV and in Fig. 11.

Practically 90 customers form eight main clusters, while the remaining four customers show specific unique characteristics among the members of the set of the 94 customers (respective individual clusters 2, 4, 9, and 10). For example, in Fig. 12 the separate customer of cluster 2 is presented, whose peak load demand occurs during the early night hours.

The obtained representative curves provide useful information about the load demand of the customers' clusters throughout the year. It is obvious that the *a priori* index of customer's activity is not representative for load curves, which is also confirmed by [1]–[5]. This cannot be generalized since it may vary from country to country and from distribution company to distribution company, depending on the respective data of customers [6]–[8]. The same process can be repeated for all other adequacy measures. The number of the clusters being used can also be selected according to the desirable precision and the relative improvement of the respective measure. It must be noticed that larger customer sets can be handled applying the same procedure, and the expected results might be better. However, only the set of 94 customers was available.

VI. PRACTICAL APPLICATION OF THE METHODOLOGY

The results of the developed methodology can be used either for each customer separately or for a set of customers. The results of the first stage, which are the typical chronological load curves of each customer, can be used as input information for

- selection of an adequate tariff by the customer or the identification of a tariff from the supplier;
- allocation of customer's bills in case of energy and power bought from more than one suppliers;
- customer's short-term and midterm load forecasting, load estimation after the application of demand-side management programs, for which both the customer and the suppliers are interested.

The results of the second stage can be used as important input information for

- adaptation of tariffs for each customer class from the suppliers [1];
- adaptation of tariffs for ancillary services of the reactive demand on behalf of the distribution or transmission operator, if the respective representative curves of reactive load are calculated;
- short-term and midterm load forecasting for the customer classes, for which the suppliers, the system operator, and the regulatory energy authority are interested.

VII. CONCLUSIONS

This paper presents an efficient two-stage pattern recognition methodology for the classification of electricity customers and for the study of electricity behavior of each customer. In the first stage, the representative daily load curves of each customer and the respective populations per each typical day are calculated. The basic contribution of the methodology is that its first stage enables the modification of the representative day, such

as the most populated day, and avoids the *a priori* definition of a single day or the "mean" day of a specific time period (as it is suggested by previously published methodologies [2]–[6]). In the second stage, the customers' classification is carried out by using the representative customer's clusters being obtained in the first stage. In both stages, the unsupervised clustering methods can be applied, such as the k-means, Kohonen adaptive vector quantization, fuzzy k-means, and hierarchical methods. The performance of these methods is evaluated by using six appropriate adequacy measures.

The additional contributions of this methodology are the following:

- formation of the typical daily load curves for each customer;
- modification of k-means and fuzzy k-means, as well as the proper parameters calibration, such as the training rate of AVQ, in order to fit the classification needs;
- comparison of the performance of the clustering algorithms in both stages by using the adequacy measures, especially the ratio of within cluster sum of squares to between cluster variation.

ACKNOWLEDGMENT

This work is dedicated to the memory of Prof. G. C. Contaxis, who supervised it until he suddenly passed away on November 1, 2004. The authors would like to thank C. Anastasopoulos, D. Voumboulakis, C. Koulopoulos, D. Lambousis, and P. Eustathiou from the Greek Public Power Cooperation (PPC) for providing the necessary data for the training of the model.

REFERENCES

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

[2] 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. 1232–1239, May 2004.

[3] G. Chicco, R. Napoli, and F. Piglione, "Application of clustering algorithms and self organising maps to classify electricity customers," in *Proc. IEEE Power Tech Conf.*, Bologna, Italy, Jun. 23–26, 2003.

[4] G. Chicco, R. Napoli, and F. Piglione, "Comparisons among clustering techniques for electricity customer classification," *IEEE Trans. Power Syst.*, vol. 21, no. 2, pp. 933–940, May 2006.

[5] V. Figueiredo, F. J. Duarte, F. Rodrigues, Z. Vale, C. Ramos, and J. B. Gouveia, "Electric customer characterization by clustering," in *Proc. ISAP*, Lemnos, Greece, Aug. 2003.

[6] D. Gerbec, S. Gasperic, and F. Gubina, "Determination and allocation of typical load profiles to the eligible consumers," in *Proc. IEEE Power Tech Conf.*, Bologna, Italy, Jun. 23–26, 2003.

[7] D. Gerbec, S. Gasperic, I. Smon, and F. Gubina, "Allocation of the load profiles to consumers using probabilistic neural networks," *IEEE Trans. Power Syst.*, vol. 20, no. 2, pp. 548–555, May 2005.

[8] D. Gerbec, S. Gasperic, I. Smon, and F. Gubina, "Determining the load profiles of consumers based on fuzzy logic and probability neural networks," *Proc. Inst. Elect. Eng., Gen., Transm., Distrib.*, vol. 151, pp. 395–400, May 2004.

[9] S. Theodoridis and K. Koutroumbas, *Pattern Recognition*, 1st ed. New York: Academic, 1999.

[10] M. Kitayama, R. Matsubara, and Y. Izui, "Application of data mining to customer profile analysis in the power electric industry," in *Proc. IEEE Power Eng. Soc. Winter Meeting*, New York, 2002.

[11] V. Figueiredo, F. Rodrigues, Z. Vale, and J. B. Gouveia, "An electric energy consumer characterization framework based on data mining techniques," *IEEE Trans. Power Syst.*, vol. 20, no. 2, pp. 596–602, May 2005.

[12] M. Petrescu and M. Scutariu, "Load diagram characterization by means of wavelet packet transformation," in *Proc. 2nd Balkan Conf.*, Belgrade, Yugoslavia, Jun. 2002, pp. 15–19.

[13] C. S. Chen, J. C. Hwang, and C. W. Huang, "Application of load survey systems to proper tariff design," *IEEE Trans. Power Syst.*, vol. 12, no. 4, pp. 1746–1751, Nov. 1997.

[14] S. Haykin, *Neural Networks, A Comprehensive Foundation*. Englewood Cliffs, NJ: Prentice-Hall, 1994.

[15] T. Kohonen, *Self-Organization and Associative Memory*, 3rd ed. New York: Springer-Verlag, 1989.

[16] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-2, pp. 224–227, Apr. 1979.

[17] D. Hand, H. Manilla, and P. Smyth, *Principles of Data Mining*. Cambridge, MA: MIT Press, 2001.

George J. Tsekouras (M'98) was born in Athens, Greece, in 1976. He received the Diploma in electrical and computer engineering, the Diploma in civil engineering, and the Ph.D. degree from the National Technical University of Athens (NTUA), Athens, Greece, in 1999, 2004, and 2006 respectively.

His research interests include power system analysis, load and energy forecasting methods, and database design.

Dr. Tsekouras is a member of the Technical Chamber of Greece.

Nikos D. Hatziargyriou (SM'90) was born in Athens, Greece, in 1954. He received the Diploma in electrical and mechanical engineering from the National Technical University of Athens (NTUA), Athens, Greece, in 1976 and the M.Sc. and Ph.D. degrees from the University of Manchester Institute of Science and Technology (UMIST), Manchester, UK, in 1979 and 1982, respectively.

He is currently a Professor at the School of Electrical and Computer Engineering of NTUA. His research interests include modeling and digital techniques for power system analysis and control.

Dr. Hatziargyriou is an SCC6 member of CIGRE and the Technical Chamber of Greece.

Evangelos N. Dialynas (SM'90) received the Diploma in electrical and mechanical engineering from the National Technical University of Athens (NTUA), Athens, Greece, in 1975 and the M.Sc. and Ph.D. degrees from the University of Manchester Institute of Science and Technology (UMIST), Manchester, UK, in 1977 and 1979, respectively.

He is currently a Professor at the School of Electrical and Computer Engineering of NTUA. His research interests include power system analysis, reliability modeling, quality assessment, and probabilistic computer applications.

Dr. Dialynas is a Distinguished member of CIGRE and a member of the Technical Chamber of Greece.