Available online at www.sciencedirect.com

ScienceDirect

Electric Power Systems Research 78 (2008) 1494–1510

ELECTRIC POWER SYSTEMS RESEARCH

www.elsevier.com/locate/epsr

A pattern recognition methodology for evaluation of load profiles and typical days of large electricity customers

G.J. Tsekouras *, P.B. Kotoulas, C.D. Tsirekis, E.N. Dialynas, N.D. Hatziargyriou

Power Systems Laboratory, School of Electrical & Computer Engineering, National Technical University of Athens, 9 Iroon Polytechniou Street, 157 73 Zografou Campus, Athens, Greece

Received 10 November 2006; received in revised form 20 June 2007; accepted 10 January 2008
Available online 4 March 2008

Abstract

This paper describes a pattern recognition methodology for the classification of the daily chronological load curves of each large electricity customer, in order to estimate his typical days and his respective representative daily load profiles. It is based on pattern recognition methods, such as *k*-means, self-organized maps (SOM), fuzzy *k*-means and hierarchical clustering, which are theoretically described and properly adapted. The parameters of each clustering method are properly selected by an optimization process, which is separately applied for each one of six adequacy measures. The results can be used for the short-term and mid-term load forecasting of each consumer, for the choice of the proper tariffs and the feasibility studies of demand side management programs. This methodology is analytically applied for one medium voltage industrial customer and synoptically for a set of medium voltage customers of the Greek power system. The results of the clustering methods are presented and discussed.

Keywords: Load profiles; Clustering algorithms; Fuzzy *k*-means; Hierarchical clustering; *k*-Means; Self-organizing map; Pattern recognition

1. Introduction

In a deregulated electricity market, each supplier wishes to identify his customers' electricity behavior accurately, in order to provide them with satisfactory services at a low cost, recovering the energy and power cost and having a fair profit. At the same time, each consumer wants to know his electricity behavior, in order to select the proper tariff or to apply energy efficiency measures successfully. Taking into consideration the demand side bidding in competitive markets [1] the accurate estimation of the next day's load profile is a fundamental requirement for each large customer, so that it can find the way to minimize its electricity bill.

During the last years, a significant research effort has been focused on load curves classification regarding the short-term load forecasting of anomalous days [2,3] and the clustering of the customers of the power systems [4–15]. The clustering methods have been used so far are the "modified follow the leader" [4–7], the self-organizing map (SOM) [2,3,5,7,8], the *k*-means [7,8], the average and Ward hierarchical methods [6,7,9] and the fuzzy *k*-means [6,7,9–11]. All the above methods generally belong to pattern recognition techniques [7–12]. Alternatively, classification problem can be solved by using data mining [13,14], wavelet packet transformation [15], frequency-domain data [16], stratified sampling [17]. For the reduction of the size of the clustering input data set Sammon map, principal component analysis and curvilinear component analysis have been proposed [7]. The respective adequacy measures that are commonly used are the mean index adequacy [4–6], the clustering dispersion indicator [4–7], the similarity matrix indicator [6], the Davies–Bouldin indicator [3,6,7,10,11], the modified Dunn index [7], the scatter index [7] and the mean square error [9–11].

The objective of this paper is to present a new methodology for the classification of the daily chronological load curves for each large electricity customer. Specifically, the load curves set of each customer is 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 or the successful application of demand side management programs. The proposed methodology compares the results

* Corresponding author. Tel.: +30 210 3479124; fax: +30 210 3479124.
E-mail addresses: tsekouras_george_j@yahoo.gr (G.J. Tsekouras), p_kwt@yahoo.gr (P.B. Kotoulas), ktsirekis@desmie.gr (C.D. Tsirekis), dialynas@power.ece.ntua.gr (E.N. Dialynas), nh@power.ece.ntua.gr (N.D. Hatziargyriou).

0378-7796/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.epsr.2008.01.010

obtained by certain clustering techniques (*k*-means with special weights initialization, mono-dimensional and bi-dimensional self-organizing maps, fuzzy *k*-means and 7 hierarchical agglomerative clustering methods) using six adequacy measures (mean square error, mean index adequacy, clustering dispersion indicator, similarity matrix indicator, Davies–Bouldin indicator, ratio of within cluster sum of squares to between cluster variation). The basic notions of this method are:

- the estimation of the typical days through the study period and the respective representative typical daily load profiles for each large electricity customer;
- the modification of the clustering techniques for this kind of classification problem, such as the appropriate weights initialization for the *k*-means and fuzzy *k*-means;
- the proper parameters calibration, such as the training rate of mono-dimensional SOM, in order to fit the classification needs;
- the comparison of the performance of the clustering algorithms for each one of the adequacy measures;
- the introduction of the ratio of within cluster sum of squares to between cluster variation, which is first presented for this kind of classification.

Finally, the results of the application of the developed methodology are analytically presented for one medium voltage industrial customer. It is also applied for a set of 94 medium voltage (MV) customers of the Greek distribution system, for which the behavior of the clustering models are synoptically registered.

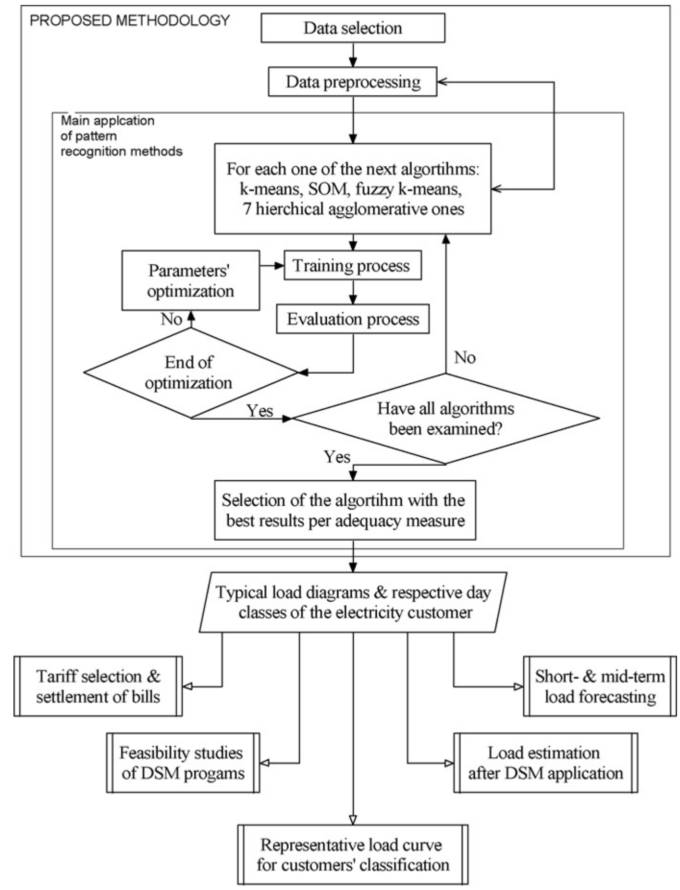

Fig. 1. Flow diagram of pattern recognition methodology for the classification of daily chronological load curves of one large electricity customer.

2. Proposed pattern recognition methodology for the classification of load curves of separate customer

The classification of daily chronological load curves of one customer is achieved by means of the pattern recognition methodology, as shown in Fig. 1. The main steps are the following:

- *Data and features selection*: using electronic meters, 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 are determined for the study period.
- *Data preprocessing*: the load diagrams of the customer are examined for normality, in order to modify or delete the values that are obviously wrong (*noise suppression*). If it is necessary, a preliminary execution of a pattern recognition algorithm is carried out, in order to track bad measurements or networks faults, which will reduce the number of the useful typical days for a constant number of clusters, if they remain uncorrected. In future, a filtering step can be added using principal component analysis, Sammon map, and curvilinear component analysis [7], for the reduction of the load diagrams dimensions.
- *Main application of pattern recognition methods*: for the load diagrams of the customer, a number of clustering algorithms (*k*-means, self-organized map, fuzzy *k*-means and hierarchical clustering) is applied. Each algorithm is trained for the set of load diagrams and evaluated according to six adequacy measures. The parameters of the algorithms are optimized, if it is necessary. The developed methodology uses the clustering methods that provide the most satisfactory results. It should be noted that conventional methods, like statistical tools, supervised techniques, etc., cannot be used, because the classification of the typical days must be already known.

The results of the developed methodology can be used for:

- the proper selection of an adequate tariff by the customer or the recommendation of a tariff from the supplier,
- the settlement of the customer's bills in the case of energy and power bought from more than one suppliers,
- the feasibility studies of the energy efficiency and demand side management measures, which are proper for the customer,
- the customer's short-term and mid-term load forecasting, load estimation after the application of demand side management programs, for which the customer as well as the suppliers are interested,
- the selection of the representative chronological load diagram of the customer 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.), which is going to be used for the customers' classification by the suppliers.

3. Mathematical modeling of clustering methods and clustering validity assessment

In the case of study of one customer's chronological typical load curves N analytical daily load diagrams are given. The main target is to find out sets of days and load patterns.

Generally N is defined as the population of the input vectors, which are going to be clustered. The $\vec{x}_\ell = (x_{\ell 1}, x_{\ell 2}, \ldots x_{\ell i}, \ldots x_{\ell d})^{\mathrm{T}}$ symbolizes the ℓ-th input vector and d is its dimension, which equals to 96 or 24, if the load measurements are taken every 15 min or every hour, respectively. The corresponding set is given by $X = \{\vec{x}_\ell : \ell = 1, \ldots, N\}$. It is worth mentioning that $x_{\ell i}$ are normalized using the upper and lower values of all elements of the original input patterns set, for the achievement of the best possible results after the application of clustering methods.

Each classification process makes a partition of the initial N input vectors to M clusters, which can be the typical days of the under study customer. The j-th cluster has a representative, which is the respective load profile and is represented by the vector $\vec{w}_j = (w_{j1}, w_{j2}, \ldots, w_{ji}, \ldots, w_{jd})^{\mathrm{T}}$ of d dimension. The last vector also expresses the cluster's center or the weight vector of neuron, if a clustering artificial neural network is used. In our case it is also called the j-th *class representative load diagram*. The corresponding set is the classes' set, which is defined by $W = \{\vec{w}_k, k = 1, \ldots, M\}$. The subset of input vectors $\vec{x}_\ell$, which belong to the j-th cluster, is Ω_j and the respective population of load diagrams is N_j.

For the study and evaluation of classification algorithms the following distances' forms are defined:

(a) the Euclidean distance between ℓ_1, ℓ_2 input vectors of the set X:

$$d(\vec{x}_{\ell_1}, \vec{x}_{\ell_2}) = \sqrt{\frac{1}{d}\sum_{i=1}^{d}(x_{\ell_1 i} - x_{\ell_2 i})^2} \tag{1}$$

(b) the distance between the representative vector $\vec{w}_j$ of j-th cluster and the subset Ω_j, calculated as the geometric mean of the Euclidean distances between $\vec{w}_j$ and each member of Ω_j:

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

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

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

3.1. 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, in order to make the load diagrams self-explanatory. The definitions of these measures are the following:

(1) *Mean square error or error function* (J) [9]:

$$J = \frac{1}{N}\sum_{\ell=1}^{N} d^2(\vec{x}_\ell, \vec{w}_{k:\vec{x}_\ell \in \Omega_k}) \tag{4}$$

(2) *Mean index adequacy* (MIA) [4], which is defined as the average of the distances between each input vector assigned to the cluster and its center:

$$\mathrm{MIA} = \sqrt{\frac{1}{M}\sum_{j=1}^{M} d^2(\vec{w}_j, \Omega_j)} \tag{5}$$

(3) *Clustering dispersion indicator* (CDI) [4], which depends on the mean infra-set distance between the input vectors in the same cluster and inversely on the infra-set distance between the class representative load curves:

$$\mathrm{CDI} = \frac{\sqrt{(1/M)\sum_{k=1}^{M}\hat{d}^2(\Omega_k)}}{\hat{d}(W)} \tag{6}$$

(4) *Similarity matrix indicator* (SMI) [6], which 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:

$$\mathrm{SMI} = \max_{p>q}\left\{\left(1 - \frac{1}{\ln[d(\vec{w}_p, \vec{w}_q)]}\right)^{-1}\right\}, \quad p, q = 1, \ldots, M \tag{7}$$

(5) *Davies–Bouldin indicator* (DBI) [18], which represents the system-wide average of the similarity measures of each cluster with its most similar cluster:

$$\mathrm{DBI} = \frac{1}{M}\sum_{k=1}^{M}\max_{p\neq q}\left\{\frac{\hat{d}(\Omega_p) + \hat{d}(\Omega_q)}{d(\vec{w}_p, \vec{w}_q)}\right\}, \quad p, q = 1, \ldots, M \tag{8}$$

(6) *Ratio of within cluster sum of squares to between cluster variation* (WCBCR) [19], which depends on the sum of the distance's square between each input vector and its cluster's representative vector, as well as the similarity of the clusters' centres:

$$\mathrm{WCBCR} = \frac{\sum_{k=1}^{M}\sum_{\vec{x}_\ell \in \Omega_k} d^2(\vec{w}_k, \vec{x}_\ell)}{\sum_{1\le q<p}^{M} d^2(\vec{w}_p, \vec{w}_q)} \tag{9}$$

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, for which the sets are empty. It is intended to minimize this number. It is noted that in Eqs. (4)–(9), M is the number of the clusters without the dead ones.

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

3.2. 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 [7,8]. In the developed algorithm the w_{ji} of the j-th center is initialized as:

$$w_{ji}^{(0)} = \frac{a + b \cdot (j-1)}{M-1} \tag{10}$$

where a and b are properly calibrated parameters. Alternatively the w_{ji} is initialized as:

$$w_{ji}^{(0)} = \frac{a_i + b_i \cdot (j-1)}{M-1} \tag{11}$$

where $a_i = \min_{\forall j}(x_{ji})$ and $b_i = \max_{\forall j}(x_{ji})$. 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 ℓ-th 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

$$\vec{w}_j^{(t+1)} = \frac{1}{N_j^{(t)}} \sum_{\vec{x}_\ell \in \Omega_j^{(t)}} \vec{x}_\ell \tag{12}$$

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 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). The best results for each adequacy measure are recorded for different pairs (a, b).

3.3. Fuzzy k-means

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

$$\sum_{j=1}^{M} u_{\ell j} = 1 \quad \& \quad 0 \le u_{\ell j} \le 1, \quad \forall j \tag{13}$$

Theoretically, the membership factor gives more flexibility in the vector's distribution. During the iterations the following objective function is minimized:

$$J_{\text{fuzzy}} = \frac{1}{N} \sum_{j=1}^{M} \sum_{\ell=1}^{N} u_{\ell j} \cdot d^2(\vec{x}_\ell, \vec{w}_j) \tag{14}$$

The membership factors and the cluster centers are calculated in each epoch as

$$u_{\ell j}^{(t+1)} = \frac{1}{\sum_{k=1}^{M} d(\vec{x}_\ell, \vec{w}_j^{(t)}) / d(\vec{x}_\ell, \vec{w}_k^{(t)})} \tag{15}$$

$$\vec{w}_j^{(t+1)} = \frac{\sum_{\ell=1}^{N} (u_{\ell j}^{(t+1)})^q \cdot \vec{x}_\ell}{\sum_{\ell=1}^{N} (u_{\ell j}^{(t+1)})^q} \tag{16}$$

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

3.4. Hierarchical agglomerative algorithms

Agglomerative algorithms are based on matrix theory [12]. 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

$$\begin{aligned} d(C_q, C_s) = {} & a_i \cdot d(C_i, C_s) + a_j \cdot d(C_j, C_s) + b \cdot d(C_i, C_j) \\ & + c \cdot |d(C_i, C_s) - d(C_j, C_s)| \end{aligned} \tag{17}$$

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 (4).

The basic algorithms, which are going to be used in our case, are

- the *single link* algorithm (SL)—it is obtained from (17) for $a_i = a_j = 0.5$, $b = 0$ and $c = -0.5$:

$$d(C_q, C_s) = \min\{d(C_i, C_s), d(C_j, C_s)\} \tag{18}$$

- the *complete link* algorithm (CL)—it is obtained from (17) for $a_i = a_j = 0.5$, $b = 0$ and $c = 0.5$:

$$d(C_q, C_s) = \max\{d(C_i, C_s), d(C_j, C_s)\} \tag{19}$$

- the *unweighted pair group method average* algorithm (UPGMA):

$$d(C_q, C_s) = \frac{n_i \cdot d(C_i, C_s) + n_j \cdot d(C_j, C_s)}{n_i + n_j} \tag{20}$$

  where n_i and n_j are the respective members' populations of clusters C_i and C_j.
- the *weighted pair group method average* algorithm (WPGMA):

$$d(C_q, C_s) = 0.5 \cdot \{d(C_i, C_s) + d(C_j, C_s)\} \tag{21}$$

- the *unweighted pair group method centroid* algorithm (UPGMC):

$$d^{(1)}(C_q, C_s) = \frac{n_i \cdot d^{(1)}(C_i, C_s) + n_j \cdot d^{(1)}(C_j, C_s)}{(n_i + n_j)} - n_i \cdot n_j \cdot \frac{d^{(1)}(C_i, C_j)}{(n_i + n_j)^2} \quad (22)$$

where $d^{(1)}(C_q, C_s) = \left\| \vec{w}_q - \vec{w}_s \right\|^2$ and $\vec{w}_q$ is the representative center of the q-th cluster (11).

- the *weighted pair group method centroid* algorithm (WPGMC):

$$d^{(1)}(C_q, C_s) = 0.5 \cdot (d^{(1)}(C_i, C_s) + d^{(1)}(C_j, C_s)) - 0.25 \cdot d^{(1)}(C_i, C_j) \quad (23)$$

- the *Ward or minimum variance* algorithm (WARD):

$$d^{(2)}(C_q, C_s) = \frac{(n_i + n_s) \cdot d^{(2)}(C_i, C_s) + (n_j + n_s) \cdot d^{(2)}(C_j, C_s) - n_s \cdot d^{(2)}(C_i, C_j)}{n_i + n_j + n_s} \quad (24)$$

where $d^{(2)}(C_i, C_j) = (n_i \cdot n_j)/(n_i + n_j) \cdot d^{(1)}(C_i, C_j)$.

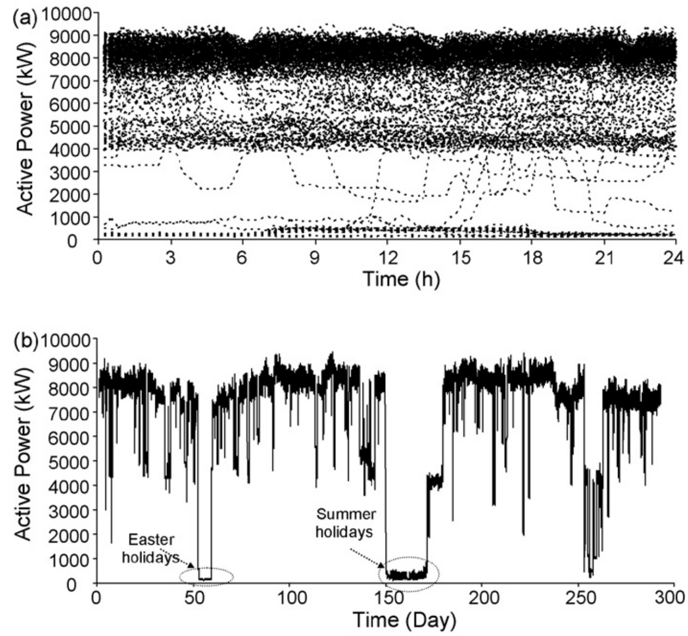

Fig. 2. Chronological 15-min load diagrams for a set of 292 training patterns for the industrial medium voltage customer (a) for each day and (b) for the time period (292 days) under study (February–November 2003).

3.5. Self-organized maps

The Kohonen SOM [20–23] is a topologically unsupervised neural network that projects a d-dimensional input data set into a reduced dimension space (usually a mono-dimensional or bi-dimensional map). It is composed of a predefined grid containing $M_1 \times M_2$ d-dimensional neurons $\vec{w}_k$, which are calculated by a competitive learning algorithm that updates not only the weights

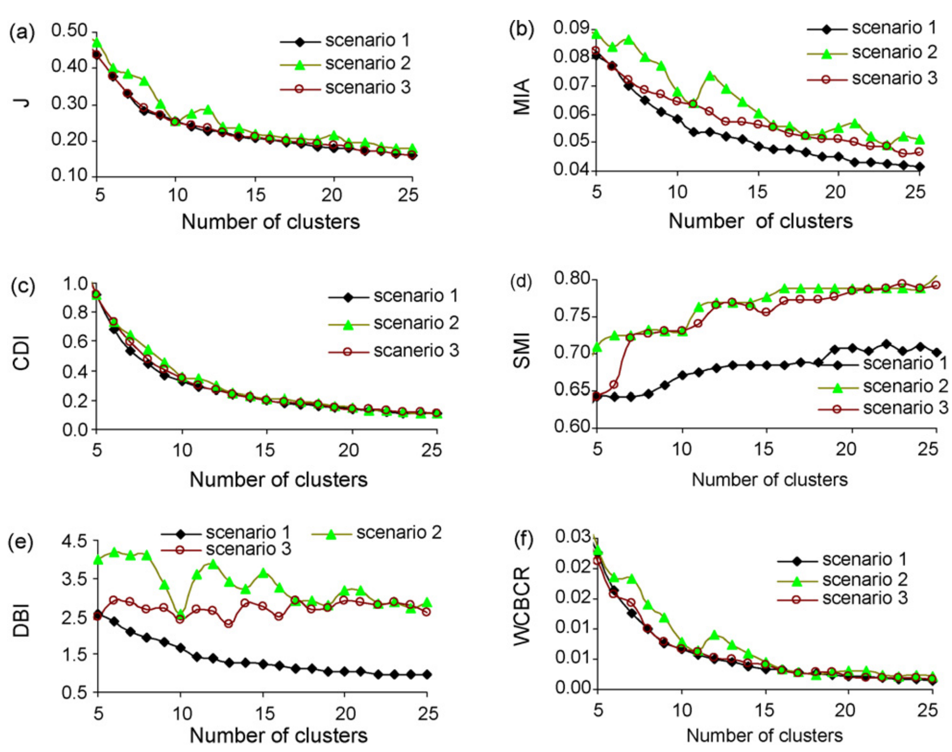

Fig. 3. Adequacy measures for the k-means method for a set of 292 training patterns for 5–25 clusters (scenario 1: proposed method – weights initialization based on Eq. (10) – , scenario 2: alternative method – weights initialization based on Eq. (11) – , scenario 3: classical method). (a) J indicator; (b) MIA indicator; (c) CDI indicator; (d) SMI indicator; (e) DBI indicator; (f) WCBCR indicator.

of the winning neuron, but also the weights of its neighbor units in inverse proportion of their distance. The neighborhood size of each neuron shrinks progressively during the training process, starting with nearly the whole map and ending with the single neuron. The training of SOM is divided to two phases:

- *rough ordering*, with high initial learning rate, large radius and small number of epochs, so that neurons are arranged into a structure which approximately displays the inherent characteristics of the input data;
- *fine tuning*, with small initial learning rate, small radius and higher number of training epochs, in order to tune the final structure of the SOM.

The transition of the rough ordering phase to the fine tuning one is happened after T_{s_0} epochs.

Once all vectors of neurons $\vec{w}_k$ have been initialized, the SOM training commences by first choosing an input vector $\vec{x}_\ell$, at t epoch, randomly from the input vectors' set. The Euclidean distances between the n-th presented input pattern $\vec{x}_\ell$ and all $\vec{w}_k$ are calculated, so as to determine the wining neuron i' that is closest to $\vec{x}_\ell$. The j-th reference vector is updated according to

$$\vec{w}_j^{(t)}(n+1) = \vec{w}_j^{(t)}(n) + \eta(t) \cdot h_{i'j}(t) \cdot (\vec{x}_\ell - \vec{w}_j^{(t)}(n)) \qquad (25)$$

where $\eta(t)$ is the learning rate according to

$$\eta(t) = \eta_0 \exp(-t/T_{\eta 0}) > \eta_{\min} \qquad (26)$$

with η_0, $\eta_{\min}$ and $T_{\eta 0}$ representing the initial value, the minimum value and the time parameter, respectively. During the rough ordering phase η_r, T_η are the initial value and the time parameter, respectively, while during the fine tuning phase the respective values are η_f, T_η. The $h_{ij}(t)$ is the neighborhood symmetrical function, that will activate the j neurons that are topologically close to the winning neuron i', according to their geometrical distance, who will learn from the same $\vec{x}_\ell$. In this case the Gauss function is proposed:

$$h_{i'j}(t) = \exp[-d_{i'j}^2/2\sigma^2(t)] \qquad (27)$$

where $d_{i'j} = \|\vec{r}_{i'} - \vec{r}_j\|$ is the respective distance between i' and j neurons, $\vec{r}_j = (x_j, y_j)$ the respective co-ordinates in the grid, $\sigma(t) = \sigma_0 \exp(-t/T_{\sigma 0})$ the decreasing neighborhood radius function where σ_0 and $T_{\sigma 0}$ are the respective initial value and time parameter of the radius, respectively.

The case studies deal with the matters of the shape of the map, the parameters' calibration and the weights' initialization. Especially, the multiplicative factors ϕ and ξ are introduced without decreasing the generalization ability of the parameters'

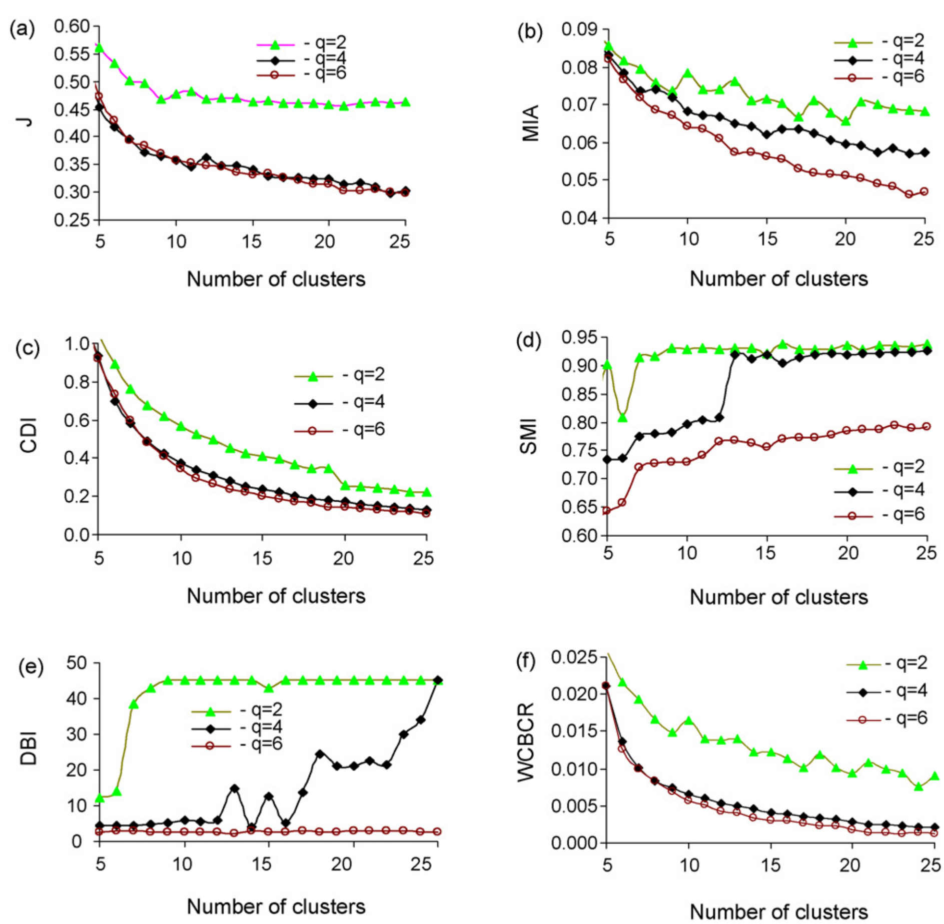

Fig. 4. Adequacy measures for the fuzzy k-means method for a set of 292 training patterns, for 5–25 clusters and $q = 2, 4, 6$. (a) J indicator; (b) MIA indicator; (c) CDI indicator; (d) SMI indicator; (e) DBI indicator; (f) WCBCR indicator.

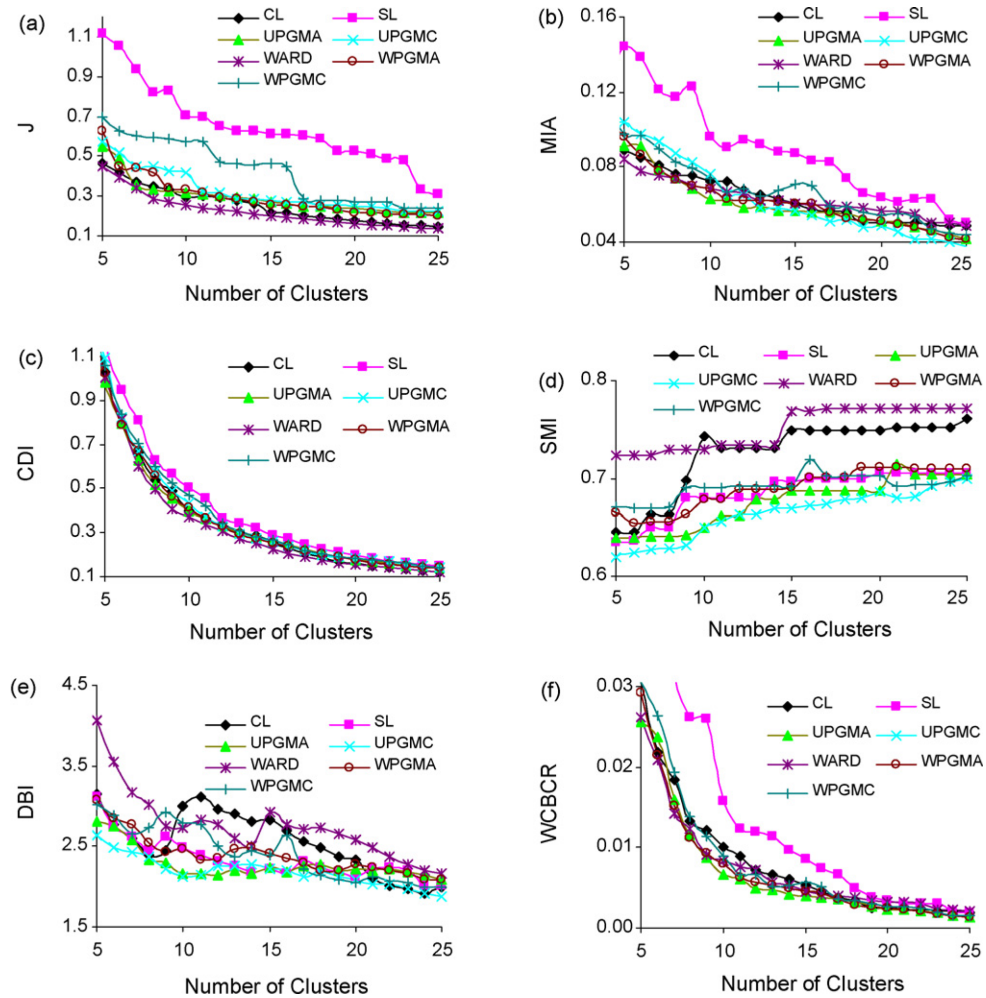

Fig. 5. Adequacy measures for the 7 hierarchical clustering algorithms for a set of 292 training patterns for 5–25 clusters. (a) *J* indicator; (b) MIA indicator; (c) CDI indicator; (d) SMI indicator; (e) DBI indicator; (f) WCBCR indicator.

calibration:

$$T_{s0} = \phi T_{\eta 0} \tag{28}$$

$$T_{\sigma 0} = \frac{\xi T_{\eta 0}}{\ln \sigma_0} \tag{29}$$

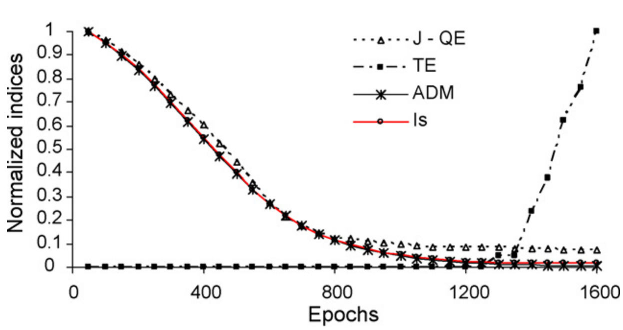

Fig. 6. Quality normalized measures of the quantization error (QE), the topographic error (TE), the average distortion measure error (ADM) and the index (Is) for the mono-dimensional SOM with 10 clusters, $\eta_r = 0.1$, $\eta_f = 0.001$, $T_{\eta 0} = 1000$, $\sigma_0 = 10$, $T_{s0} = T_{\eta 0}$, $T_{\sigma 0} = T_{\eta 0}/\ln \sigma_0$ in the case of a set of 292 training patterns of the industrial customer under study.

It is mentioned, that the quality measures of the optimum SOM are based on the quantization error given by (4), the topographic error and the average distortion measure. The topographic error measures the distortion of the map as the percentage of input

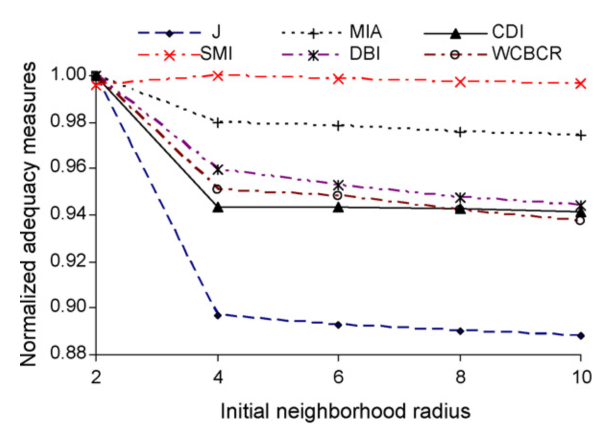

Fig. 7. Normalized adequacy measures with respect to the initial radius σ_0 for the mono-dimensional SOM with 10 clusters, $\eta_r = 0.1$, $\eta_f = 0.001$, $T_{\eta 0} = 1000$, $T_{s0} = T_{\eta 0}$, $T_{\sigma 0} = T_{\eta 0}/\ln \sigma_0$ in the case of a set of 292 training patterns of the industrial customer under study.

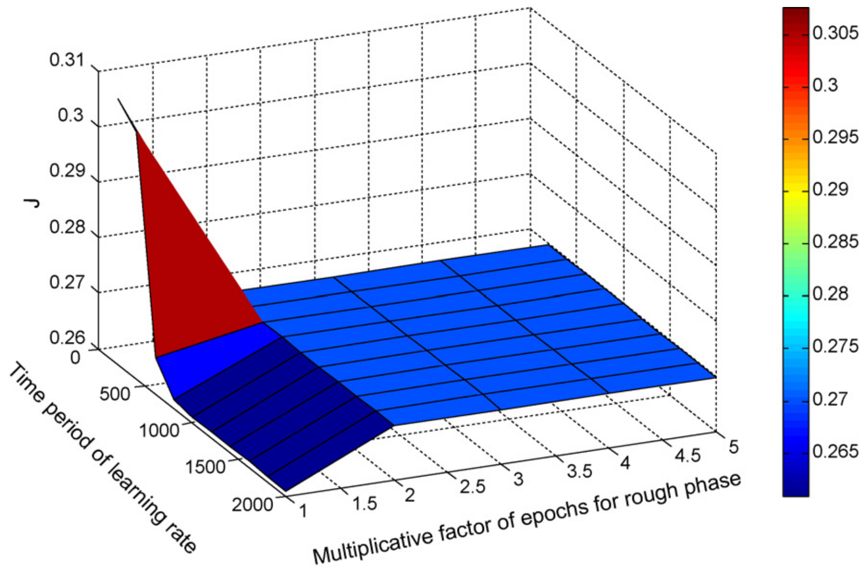

Fig. 8. Adequacy measure J with respect to $\phi = \{1, 2, 3, 4, 5\}$ and $T_{\eta 0} = \{200, 400, \ldots, 2000\}$ for the mono-dimensional SOM with 10 clusters, $\eta_r = 0.1$, $\eta_f = 0.001$, $\sigma_0 = 10$, $T_{\sigma_0} = T_{\eta 0}/\ln \sigma_0$ in the case of a set of 292 training patterns for the industrial customer under study.

vectors for which the first i'_1 and second i'_2 winning neuron are not neighboring map units:

$$\mathrm{TE} = \frac{\sum_{\ell=1}^{N} \mathrm{neighb}(i'_1, i'_2)}{N} \tag{30}$$

where for each input vector, $\mathrm{neighb}(i'_1, i'_2)$ equals to 1, if i'_1 and i'_2 neurons are not neighbors, either 0. The average distortion measure is given for the t epoch by

$$\mathrm{ADM}(t) = \frac{\sum_{\ell=1}^{N} \sum_{j=1}^{M} h_{i' \to \vec{x}_\ell, j}(t) \cdot d^2(\vec{x}_\ell, \vec{w}_j)}{N} \tag{31}$$

The supposed best trained SOM is the one trained with a number of epochs t for which the following index Is gets the minimum value [22]:

$$\mathrm{Is}(t) = J(t) + \mathrm{ADM}(t) + \mathrm{TE}(t) \tag{32}$$

4. Application of the proposed methodology to a medium voltage customer

The developed methodology was analytically applied on one medium voltage industrial paper mill customer of the Greek distribution system. The data used are 15 min load values for a period of 10 months in 2003. The respective set of the daily chronological curves has 301 members. Nine curves were rejected through data pre-processing, while the remaining 292 diagrams were used by the aforementioned clustering methods. The last diagrams are registered in Fig. 2, in which the load variability is also presented. The load behavior is significantly decreased during holiday time. The mean load demand

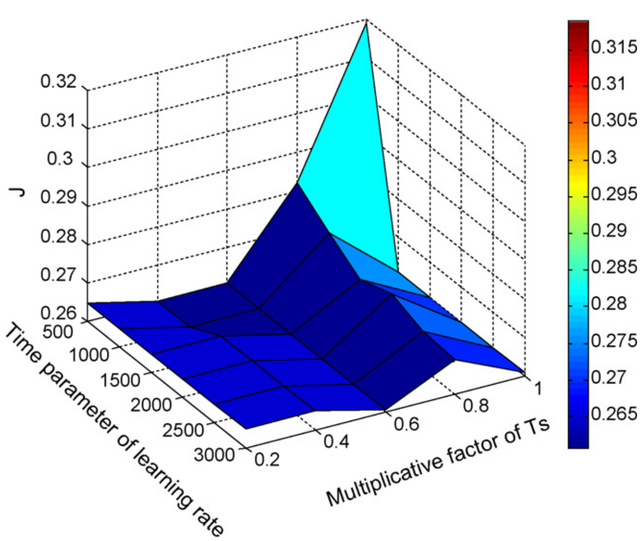

Fig. 9. Adequacy measure J with respect to $\xi = \{0.2, 0.4, \ldots, 1.0\}$ and $T_{\eta 0} = \{500, 1000, \ldots, 3000\}$ for the mono-dimensional SOM with 10 clusters, $\eta_r = 0.1$, $\eta_f = 0.001$, $\sigma_0 = 10$, $T_{\sigma_0} = T_{\eta 0}/\ln \sigma_0$ in the case of a set of 292 training patterns for the industrial customer under study.

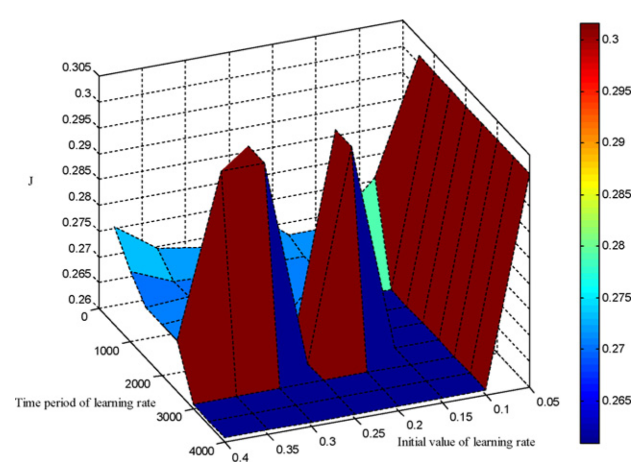

Fig. 10. Adequacy measure J with respect to $\eta_r = (0.05, 0.10, \ldots, 0.4)$ and $T_{\eta 0} = \{500, 1000, \ldots, 4000\}$ for the mono-dimensional SOM with 10 clusters, $\eta_f = 0.001$, $\sigma_0 = 10$, $T_{s_0} = T_{\eta 0}$, $T_{\sigma_0} = T_{\eta 0}/\ln \sigma_0$ in the case of a set of 292 training patterns for the industrial customer under study.

Table 1
Quality indices for different cases of bi-dimensional SOM

2D SOM-neurons population	Arrangement–weights initialization	Total epochs, t	$I(t)$	ADM(t)	TE(t)	$J(t)$	Calibration of T_η–ξ–ϕ–η_r–η_f–σ_0
9 × 9 = 81	Rect.–(2)	3600	0.2832	0.08853	0.10616	0.08853	1500–0.4–2–0.15–0.001–9
9 × 9 = 81	Hex.–(2)	2400	0.3245	0.08372	0.15753	0.08372	1000–1.0–2–0.20–0.001–9
10 × 10 = 100	Rect.–(2)	3600	0.2208	0.06757	0.08562	0.06757	1500–0.6–2–0.40–0.001–10
10 × 10 = 100	Hex.–(2)	2400	0.2354	0.07147	0.09247	0.07147	1000–1.0–2–0.25–0.001–10
12 × 12 = 144	Rect.–(2)	1100	0.2371	0.05176	0.13356	0.05176	500–1.0–2–0.15–0.001–12
12 × 12 = 144	Hex.–(2)	4400	0.2679	0.05003	0.16781	0.05003	2000–1.0–2–0.05–0.001–12
14 × 14 = 196	Rect.–(2)	3600	0.2306	0.04167	0.14726	0.04167	1500–1.0–2–0.05–0.001–14
14 × 14 = 196	Hex.–(2)	2200	0.2245	0.03521	0.15410	0.03521	500–1.0–2–0.05–0.001–16
16 × 16 = 256	Rect.–(2)	2200	0.1356	0.02840	0.07877	0.02840	1000–1.0–2–0.10–0.001–16
16 × 16 = 256	Hex.–(2)	3300	0.1678	0.02909	0.10959	0.02909	1500–1.0–2–0.10–0.001–16
19 × 19 = 361	Rect.–(2)	4400	0.0970	0.01254	0.07192	0.01254	2000–1.0–2–0.15–0.001–19
19 × 19 = 361	Hex.–(2)	4400	0.1267	0.01538	0.09589	0.01538	2000–1.0–2–0.05–0.001–19
55 × 2 = 110	Rect.–(2)	1200	0.3504	0.05532	0.23973	0.05532	500–1.0–2–0.15–0.001–55
55 × 2 = 110	Rect.–(1)	1200	0.3503	0.05532	0.23973	0.05532	500–1.0–2–0.15–0.001–55
82 × 3 = 246	Rect.–(2)	1100	0.2040	0.01982	0.16438	0.01982	500–1.0–2–0.30–0.001–82

is 6656 kW and the peak load demand is 9469 kW during the period under study.

The main goal of the application of this methodology is the representation of the load behavior of the customer with typical daily load chronological diagrams. This is achieved through the following steps:

- The calibration of the parameters of each clustering method is realized for every adequacy measure separately and the performance for different number of clusters is registered.
- The clustering models are compared for the six adequacy measures, the behavior of these measures is studied and the appropriate number of the clusters is defined.

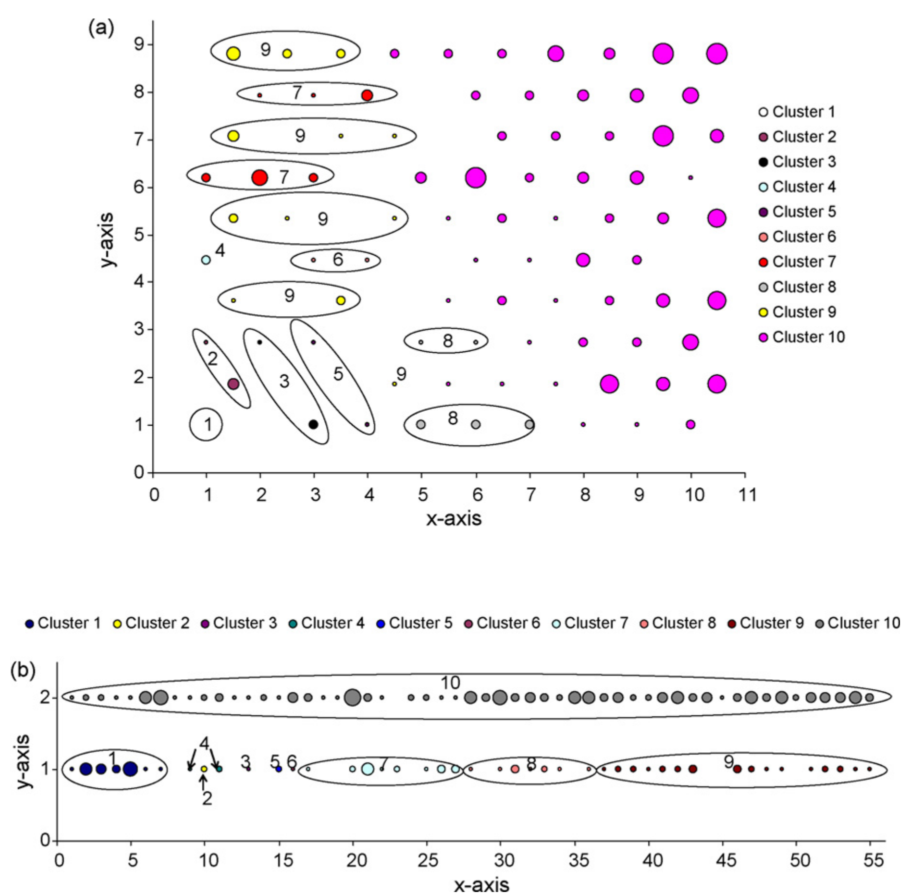

Fig. 11. Different cases of bi-dimensional SOM after the application of the proposed k-means method at the neurons of SOM (a) 10 × 10 with hexagonal arrangement and (b) 55 × 2 with rectangular arrangement.

Table 2
Adequacy indices for 10 clusters—typical load chronological curves of the industrial customer using proposed *k*-means method at the second classification level for different cases of bi-dimensional SOM

2D SOM-neurons population	Arrangement–weights initialization	Adequacy measure					
		J	MIA	CDI	SMI	DBI	WCBCR ($\times 10^{-3}$)
$9 \times 9 = 81$	Rect.–(2)	0.309430	0.064420	0.365331	0.668309	2.01767	8.4570
$9 \times 9 = 81$	Hex.–(2)	0.284911	0.064730	0.358870	0.666259	1.91128	9.1605
$10 \times 10 = 100$	Rect.–(2)	0.269351	0.066746	0.358512	0.661059	1.76758	8.7586
$10 \times 10 = 100$	Hex.–(2)	0.285281	0.062488	0.369192	0.682404	1.81237	8.4091
$12 \times 12 = 144$	Rect.–(2)	0.267056	0.065584	0.353097	0.671649	1.74857	9.2029
$12 \times 12 = 144$	Hex.–(2)	0.268810	0.070409	0.351914	0.661194	1.85451	9.4168
$14 \times 14 = 196$	Rect.–(2)	0.272213	0.066418	0.360227	0.662015	1.68854	8.8566
$14 \times 14 = 196$	Hex.–(2)	0.273781	0.069001	0.361199	0.666046	1.75621	10.0756
$16 \times 16 = 256$	Rect.–(2)	0.267521	0.066970	0.349423	0.669469	1.88224	9.2304
$16 \times 16 = 256$	Hex.–(2)	0.268528	0.068710	0.364127	0.660511	1.69430	9.4152
$19 \times 19 = 361$	Rect.–(2)	0.266128	0.065389	0.351903	0.682931	1.85560	8.5386
$19 \times 19 = 361$	Hex.–(2)	0.267087	0.067618	0.343808	0.660950	1.70171	8.8486
$55 \times 2 = 110$	Rect.–(2)	0.262634	0.060581	0.345677	0.654891	1.68728	7.7872
$82 \times 3 = 246$	Rect.–(2)	0.258002	0.063284	0.334516	0.681566	1.75790	8.1426

- The representative daily load chronological diagrams of the customer are calculated for the best clustering techniques and the proposed number of clusters.

4.1. Application of the k-means

The proposed model of the *k*-means method (*k*-means-scenario 1 with the weights initialization based on Eq. (10)) was executed 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 cluster, 1332 different pairs (a, b) were checked. The best results for the 6 adequacy measures did not refer to the same pair (a, b). The second model of the *k*-means method (*k*-means-scenario 2) was based on Eq. (11) for the weights initialization. The third model (*k*-means-scenario 3) was the classical one with the random choice of the input vectors during the centers' initialization. For the classical *k*-means model, 100 executions were carried out and the best results for each index were registered. In Fig. 3, it is obvious that the proposed *k*-means is superior to the other two scenarios of *k*-means. The superiority of the proposed model applies in all cases of neurons.

A second advantage comprise the convergence to the same results for the respective pairs (a, b), which cannot be reached using the classical model.

4.2. 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 shown in Fig. 4, where the six adequacy measures are pre-

Table 3
Comparison of the best clustering models for 10 clusters for a medium voltage customer

Methods–parameters	Adequacy measure					
	J	MIA	CDI	SMI	DBI	WCBCR
Proposed *k*-means (scenario 1)	0.2527	0.05828	0.3239	0.6711	1.6515	0.006679
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
k-Means (scenario 2)	0.2537	0.06782	0.3601	0.7311	2.5603	0.007760
Classical *k*-means (scenario 3)	0.2538	0.06435	0.3419	0.7306	2.4173	0.006716
Fuzzy *k*-means ($q = 6$)	0.3575	0.07144	0.3697	0.7635	3.2559	0.007153
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
CL	0.2973	0.07271	0.3977	0.7427	2.9928	0.010052
SL	0.7027	0.09644	0.5049	0.6798	2.4855	0.015696
UPGMA	0.3127	0.06297	0.4008	0.6494	2.1714	0.006684
UPGMC	0.4147	0.07656	0.4346	0.6494	2.1198	0.009908
WARD	0.2538	0.06804	0.3723	0.7296	2.7334	0.008399
WPGMA	0.3296	0.06807	0.4112	0.6781	2.4764	0.007991
WPGMC	0.5747	0.07386	0.4665	0.6900	2.7903	0.008884
Mono-dimensional SOM	0.2607	0.07189	0.3903	0.7588	3.8325	0.009631
Bi-dimensional SOM 55×2 using proposed *k*-means for classification in a second level	0.2623	0.06059	0.3456	0.6549	1.6873	0.007787
a parameter of *k*-means	0.15	0.24	0.36	0.44	0.22	0.22
b parameter of *k*-means	0.69	0.28	0.51	0.11	0.30	0.30

sented for different number of clusters and for three cases of $q = \{2, 4, 6\}$. It is noted that the initialization of the respective weights is similar to the proposed *k*-means.

4.3. Application of hierarchical agglomerative algorithms

In the case of the 7 hierarchical models the best results are given by the WARD model for *J* and CDI adequacy measures and by the UPGMA model for MIA, SMI, WCBCR indicators. For the *Davies–Bouldin* indicator there are significant variances, according to Fig. 5.

4.4. Application of mono-dimensional self-organizing map

Although the SOM algorithm is theoretically well defined, there are several issues that need to be solved for the effective

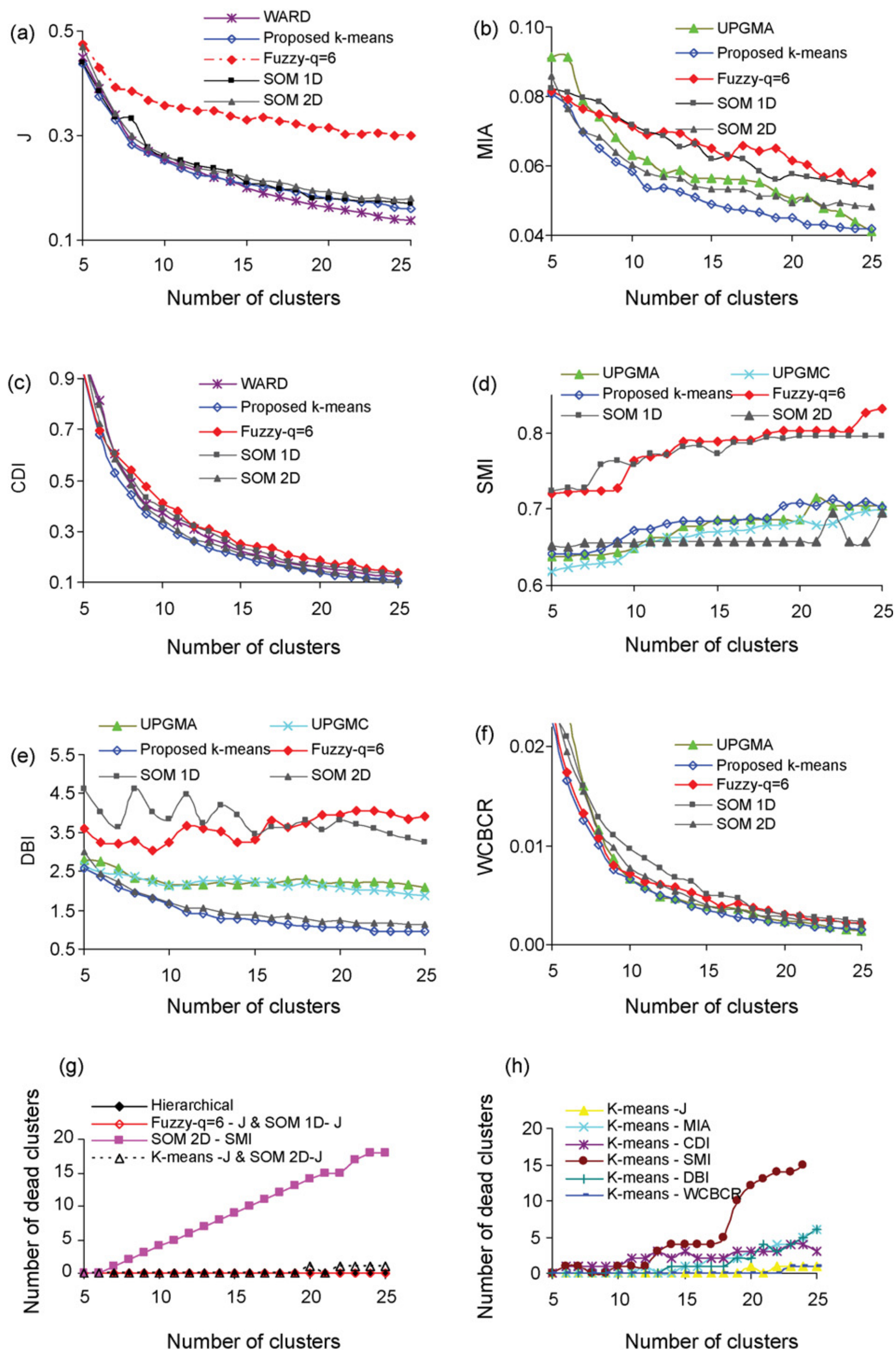

Fig. 12. The best results of each clustering method for the set of 292 training patterns of a medium voltage industrial customer for 5–25 clusters. (a) *J* indicator; (b) MIA indicator; (c) CDI indicator; (d) SMI indicator; (e) DBI indicator; (f) WCBCR indicator; (g) dead clusters for the basic clustering methods; (h) dead clusters for proposed *k*-means method.

training of SOM. The major problems are:

- to stop the training process of the optimum SOM. In this case the target was to minimize the index Is (Eq. (32)), which combines the quality measures of the quantization error given by Eq. (4), the topographic error given by Eq. (30) and the average distortion measure error given by Eq. (31). In Fig. 6, the normalized values of these four indices are registered for the case of a 10×1 SOM for the chronological load curves of the industrial customer under study;
- the proper initial value of the neighborhood radius σ_0. The radius follows the decreasing power neighborhood radius function $\sigma(t) = \sigma_0 \exp(-t/T_{\sigma 0})$, which has the advantage to act on all neurons of the map with decreasing weights according to the respective distances of the winning neuron. On the contrary, the linear radius function does not change the weights of those neurons with distances from the winning neuron bigger than $\sigma(t)$. The computational time of the last one is significantly smaller than the power function. In Fig. 7 the effects of the initial radius σ_0 to the adequacy measures are registered;
- the proper values of (a) the multiplicative factor ϕ between T_{s_0} (epochs of the rough ordering phase) and $T_{\eta 0}$ (time parameter of learning rate), (b) the multiplicative factor ξ between $T_{\sigma 0}$ (time parameter of neighborhood radius) and $T_{\eta 0}$, (c) the learning rate η_r during the rough ordering phase, (d) the learning rate η_f during the fine tuning phase. In Figs. 8–10 the adequacy measure J with respect to ϕ and $T_{\eta 0}$, with respect to ξ and $T_{\eta 0}$, and with respect to η_r and $T_{\eta 0}$ are presented, respectively, as indicative examples. From the results of the performed study, it is derived that the proper value of the parameter η_f must be smaller than 20% of the initial value of the learning rate η_r and between 10^{-3} and 10^{-4};
- the initialization of the weights of the neurons. Three cases were examined: (a) $w_{ki} = 0.5, \forall k, i$, (b) the random initialization of each neuron's weight, (c) the random choice of the input vectors for each neuron. The best training behavior was presented in case (a).

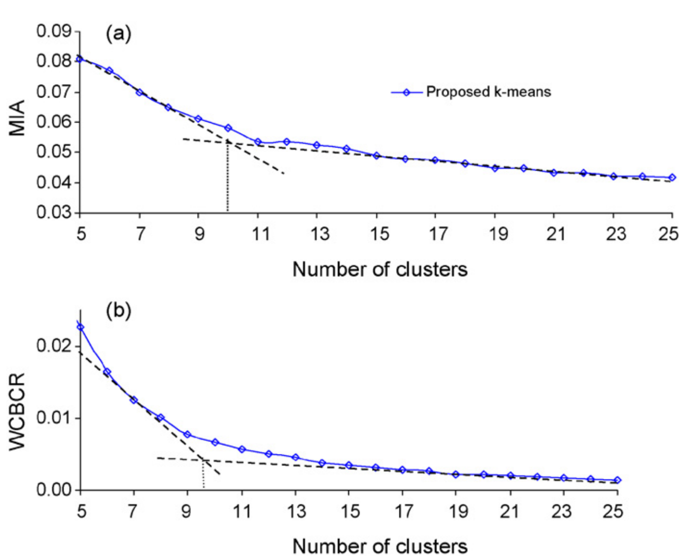

Fig. 13. Indicative estimation of the necessary clusters for the typical load daily chronological curves of a medium voltage industrial customer, for the (a) MIA and (b) WCBCR adequacy indicators.

Table 4
Results of the proposed k-means model with optimization to WCBCR adequacy measure for 10 clusters for the medium voltage industrial customer

Load cluster	Day (1 for Monday, 2 for Tuesday, 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

The optimization process for the mono-dimensional SOM parameters is repeated for any population of clusters.

4.5. Application of bi-dimensional self-organizing map

In the case of the bi-dimensional SOM the shape, the population of neurons and their respective arrangement are issues to be solved beyond the optimization of parameters, which is considered during the training process of the mono-dimensional SOM.

The rectangular shape of the map is defined with rectangular or hexagonal arrangement of neurons. The population of last ones is recommended to be $5 \times \sqrt{N}$ to $20 \times \sqrt{N}$ [4,22,23]. The height/width ratio M_1/M_2 of the rectangular grid can be calculated as the ratio between the two major eigenvalues of the covariance matrix of the input vectors set. Alternatively, the grid can be a square with properly selected sizes. The initialization of the neurons can be a linear combination of the respective eigen-

Table 5
Results of the proposed k-means model with optimization to MIA adequacy measure for 10 clusters for the medium voltage industrial customer

Load cluster	Day (1 for Monday, 2 for Tuesday, 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	0	1	0	2
3	1	0	0	0	2	0	1	4
4	0	0	1	0	0	0	0	1
5	0	1	0	0	0	0	0	1
6	0	1	0	0	0	0	0	1
7	0	0	1	1	0	0	0	2
8	4	3	3	3	3	2	2	20
9	8	5	6	7	6	4	2	38
10	23	27	26	27	27	32	34	196
	Total number of days under study							292

vectors of the two major eigenvalues (scenario of initialization 1) or can be equal to 0.5 (scenario of initialization 2). In the case of the industrial customer, a set of 292 vectors was given. The map can have 85 ($\cong 5 \times \sqrt{292}$) to 342 ($\cong 20 \times \sqrt{292}$) neurons. The respective square maps can be 9×9 to 19×19. Using the ratio between the two major eigenvalues, the respective ratio is 27.31 (=4.399/0.161) and the proposed grids can be 55×2 and 82×3. In Table 1 the quality indices are presented for different grids, arrangements of neurons, weights' initialization. The best result for the index Is is given for the square grid 19×19.

The kind of the arrangement and the weights initialization do not affect the respective results significantly. Practically, the clusters of the bi-dimensional map cannot be directly exploited because of the size and the location of the neurons into the grid, as shown in Fig. 11. This problem is solved through the application of a basic classification method – like the proposed k-means – for the neurons of the bi-dimensional SOM [4]. The adequacy measures are calculated using the load daily chronological curves of the neurons which form the respective clusters of the basic classification method. In Table 2, the adequacy measures of the aforementioned maps are presented, using the proposed k-means method for 10 final clusters.

For the industrial customer, the best results of the application of the k-means method to the neurons of the SOM are given for

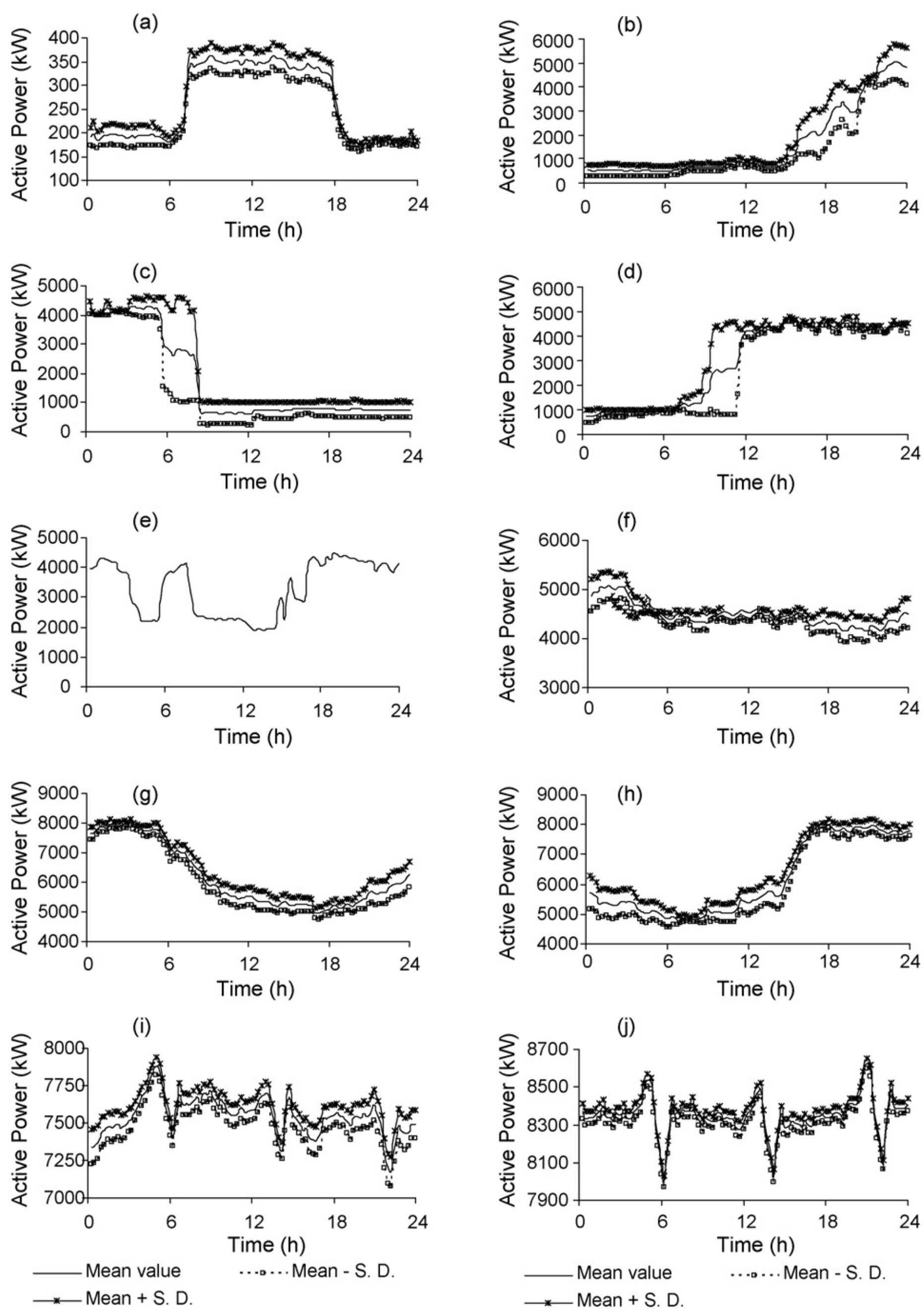

Fig. 14. Typical daily chronological load curves for the medium voltage industrial customer using proposed k-means model with optimization to WCBCR adequacy measure. (a) Cluster 1; (b) cluster 2; (c) cluster 3; (d) cluster 4; (e) cluster 5; (f) cluster 6; (g) cluster 7; (h) cluster 8; (i) cluster 9; (j) cluster 10.

the maps with the ratio between the two major eigenvalues of the covariance matrix of the input vectors set (see Table 2). The respective clusters are also more compact than the ones of the square maps, as it can be seen in Fig. 11.

4.6. Comparison of clustering models & adequacy indicators

In Fig. 12, the best results for each clustering method (proposed *k*-means, fuzzy *k*-means, self-organized maps and hierarchical algorithms) are registered. The proposed *k*-means model has the smallest values for the MIA, CDI, DBI and WCBCR indicators. The WARD algorithm presents the best behavior for the mean square error *J*, the unweighted pair group method average algorithm (UPGMA) and the bi-dimensional SOM (with the application of the proposed *k*-means at the second level) for the SMI indicator. The proposed *k*-means model has similar behavior to the WARD algorithm for the *J* indicator and to the UPGMA algorithm for the WCBCR. All indicators – except DBI – exhibit improved performance, as the number of clusters is increased.

Observing the number of dead clusters for the proposed *k*-means model (Fig. 12h), it is obvious that the use of WCBCR indicator is slightly superior to the use of MIA and *J* indica-

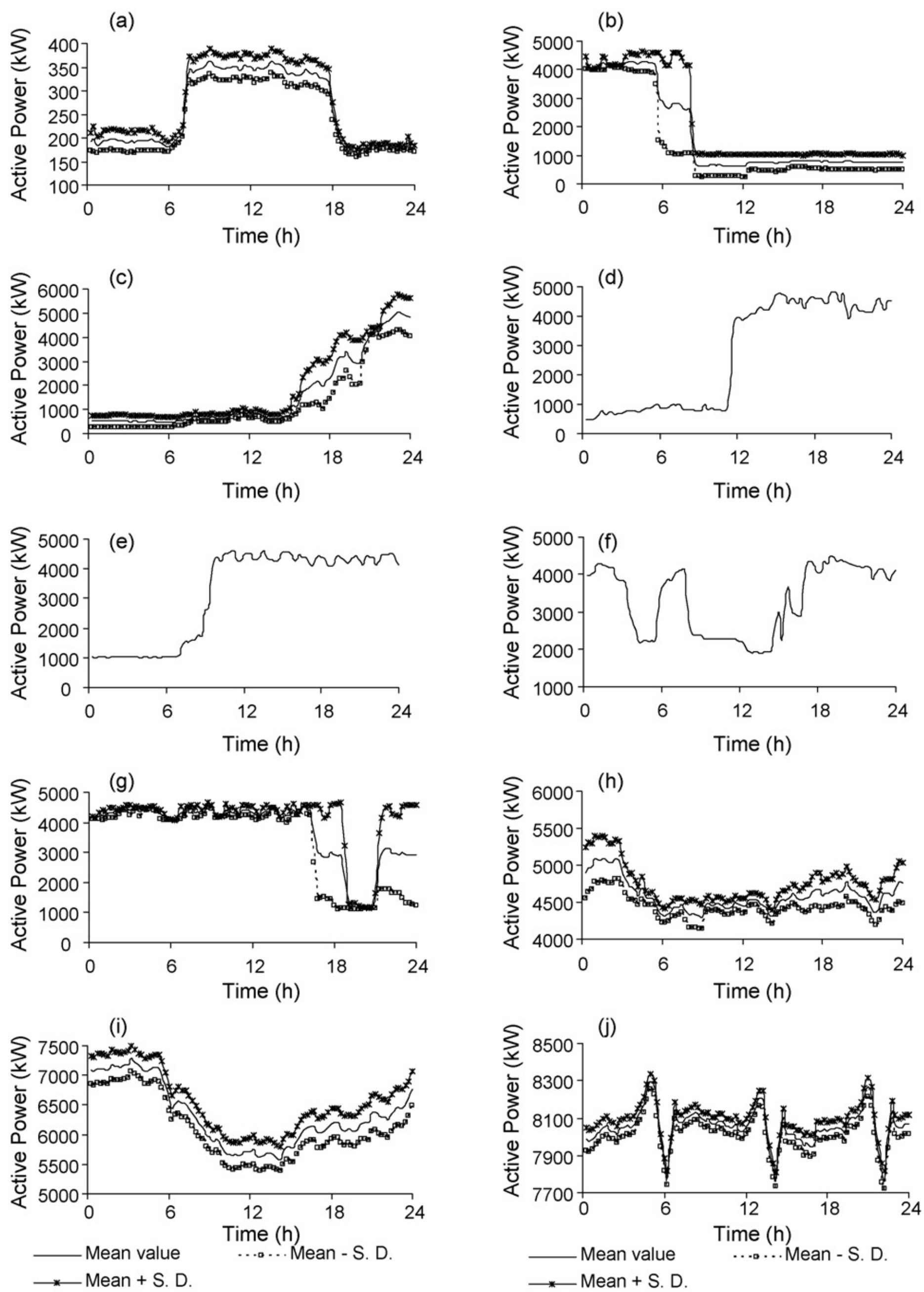

Fig. 15. Typical daily chronological load curves for the medium voltage industrial customer using proposed *k*-means model with optimization to MIA adequacy measure. (a) Cluster 1; (b) cluster 2; (c) cluster 3; (d) cluster 4; (e) cluster 5; (f) cluster 6; (g) cluster 7; (h) cluster 8; (i) cluster 9; (j) cluster 10.

tors. It is also noted, that the basic theoretical advantage of the WCBCR indicator 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. The behavior of DBI and SMI indicators for mono-dimensional SOM and fuzzy k-means emerges a significant variability. For the above reasons the proposed indicators are MIA and WCBCR.

The improvement of the adequacy indicators is significant for the first 10 clusters. After this point, the behavior of the most indicators is gradually stabilized. It can also be estimated graphically by using the rule of the "knee" [10,11], as shown in Fig. 13.

In Table 3 the results of the best clustering methods are presented for 10 clusters, which is the finally proposed size of the typical days for that customer. The proposed k-means method gives the best results for the J, MIA, CDI, DBI and WCBCR indicators (for different pairs of (a, b)), while the bi-dimensional self-organized map using proposed k-means for classification in a second level for SMI indicator.

After having taken into consideration that the analogy of the computational training time for the under study methods is 0.05:1:24:36:50 (hierarchical: proposed k-means: mono-dimensional SOM: fuzzy k-means for $q = 6$: bi-dimensional SOM), the use of the hierarchical and k-means models is proposed. It is mentioned that the necessary computational training time for the proposed k-means method is approximately 1 h for Pentium 4, 1.7 GHz, 768 MB.

4.7. Representative daily load chronological diagrams of customer

The results of the respective clustering for 10 clusters using the proposed k-means model with the optimization of the WCBCR and MIA indicators are presented in Tables 4 and 5 and in Figs. 14 and 15, respectively. This number of clusters is qualitatively satisfied.

The retailer and the head engineer of the under study industry can observe the customer's daily demand behavior during the year based on the respective load curves. Specifically, in the case of the proposed k-means model with the optimization of the WCBCR indicator, the cluster 1 represents holidays, the clusters 2 and 4 the days of the re-operation of the industry, the cluster 3 the days of stopping the operation of the industry, the cluster 5 a day with partial internal power fault, the cluster 6 the workdays with one of the two lines for production in operation, the clusters 7 and 8 the workdays for which one of the two lines for production is out of operation for few hours, the clusters 9 and 10 the usual workdays, where every 8 h there is a small variance because of the workers' change.

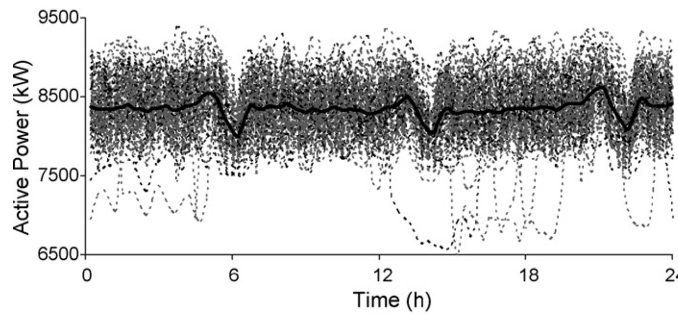

Fig. 16. Daily chronological load curve of cluster 10 for the MV industrial customer (bold line) along with its 120 clustered measured curves (thin lines) using the proposed k-means clustering method and the WCBCR indicator.

In the case of the proposed k-means model with the optimization of the MIA indicator, the cluster 1 also represents holidays, while the clusters 2–7 represent the days with special characteristics (the re-operation of the industry, stopping the operation of the industry, power fault, etc.). The cluster 8 represents the workdays with one of the two lines for production in operation, the cluster 9 the workdays for which one of the two lines for production is out of operation for few hours, the cluster 10 the usual workdays.

Using the MIA indicator emphasis is given to the days with special characteristics, while the usual workdays are united to one cluster. The final selection of the indicator, for which the optimization is carried out, depends on the user. In this case, the use of the WCBCR indicator is proposed, for the derivation of

Table 6
Comparison of the clustering models for the set of 94 MV customers for 10 clusters

Methods	Adequacy measure					
	J	MIA	CDI	SMI	DBI	WCBCR
Proposed k-means (scenario 1)	6	29	85	31	32	37
k-Means (scenario 2)	0	0	0	0	0	0
Classical k-means (scenario 3)	0	0	0	0	0	0
Fuzzy k-means ($q = 6$)	0	0	0	0	3	6
CL	0	1	0	0	1	0
SL	0	5	0	2	5	4
UPGMA	0	16	0	4	13	13
UPGMC	0	19	0	26	13	18
WARD	18	0	0	0	0	1
WPGMA	0	13	0	0	5	8
WPGMC	0	7	0	3	4	7
Mono-dimensional SOM	5	1	0	0	0	0
Bi-dimensional SOM using proposed k-means for classification in a second level	65	3	9	28	18	0

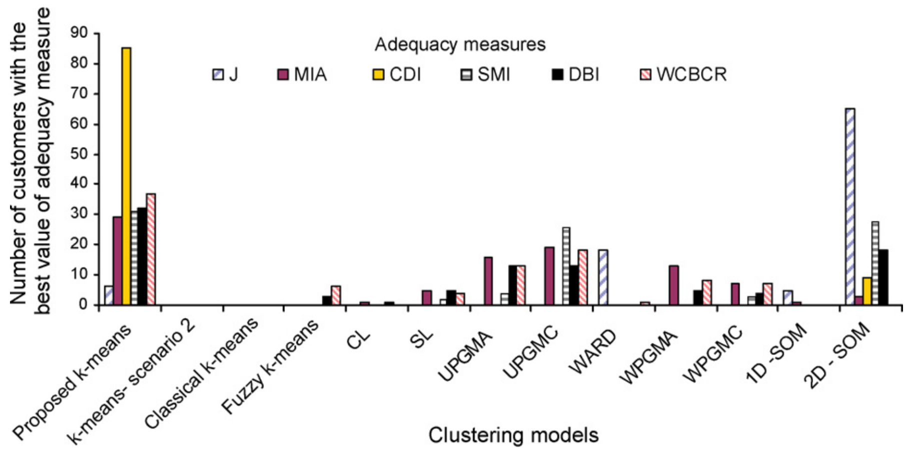

Fig. 17. Population of customers with the best value of adequacy measure, with respect to different clustering models for the set of the 94 medium voltage customers of the Greek power distribution system.

an analytical description of the load demand among the usual workdays.

The daily load diagrams are well identified using the *k*-means clustering method and the WCBCR indicator, as it is indicatively presented in Fig. 16, where 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.

5. Application of the proposed methodology to a set of medium voltage customers

The same process was repeated for 93 more customers, with load curves qualitatively described by using 8–12 clusters for each customer. The scope of this application is the representation of the comparison of the clustering algorithms and the adequacy measures for more than one customer.

The performance of these methods is presented in Table 6 and in Fig. 17, through the indication of the number of customers which achieve the best value of adequacy measure.

In Fig. 17 the comparison of the algorithms showed that the developed *k*-means method achieves a better performance for MIA, CDI, SMI, DBI and WCBCR measures, the bi-dimensional SOM model using proposed *k*-means for classification in a second level for *J* measure. It can be noticed that the other two *k*-means models show the worst performance in adequacy measures.

6. Conclusions

This paper presents an efficient pattern recognition methodology for the study of the electricity behavior of one separate customer. Any unsupervised clustering methods, such as the *k*-means, self-organized maps, fuzzy *k*-means and hierarchical methods, can be applied. The performance of these methods is evaluated by six appropriate adequacy measures. The representative daily load diagrams of each customer and the respective populations per each typical day are calculated. This information is valuable for the suppliers or for the customers, because it allows the load forecasting, the tariffs selection and the application of demand side management or energy efficiency programs.

From the respective application for a set of medium voltage customers it is concluded that 8–12 clusters are necessary for the satisfactory description of the daily load curves of each customer during 1 year.

Acknowledgements

This paper is dedicated to the memory of Prof. George C. Contaxis, who supervised it until he suddenly passed away on 1st November 2004. The authors would like to thank C. Anastasopoulos, D. Voumboulakis, C. Koulopoulos, D. Lambousis and P. Eustathiou from PPC for providing all the necessary data for the training of the model.

References

[1] Task VIII of International Energy Agency, Demand-Side Bidding in a Competitive Electricity Market, Research Report No. 1, Ver. 2, Market participants' views towards and experiences with Demand Side Bidding, January, 2002.

[2] R. Lamedica, A. Prudenzi, M. Sforna, M. Caciotta, V.O. Cencelli, A neural network based technique for short-term forecasting of anomalous load periods, IEEE Trans. Power Syst. 11 (1996) 1749–1756.

[3] M. Beccali, M. Cellura, V. Lo Brano, A. Marvuglia, Forecasting daily urban electric load profiles using artificial neural networks, Energy Conv. Manage. 45 (2004) 2879–2900.

[4] G. Chicco, R. Napoli, P. Postolache, M. Scutariu, C. Toader, Customer characterization for improving the tariff offer, IEEE Trans. Power Syst. 18 (2003) 381–387.

[5] G. Chicco, R. Napoli, F. Piglione, P. Postolache, M. Scutariu, C. Toader, Load pattern-based classification of electricity customers, IEEE Trans. Power Syst. 19 (2004) 1232–1239.

[6] G. Chicco, R. Napoli, F. Piglione, Application of clustering algorithms and self organizing maps to classify electricity customers, Presented at the IEEE Power Tech Conference, Bologna, Italy, June 23–26, 2003.

[7] G. Chicco, R. Napoli, F. Piglione, Comparisons among clustering techniques for electricity customer classification, IEEE Trans. Power Syst. 21 (2006) 933–940.

[8] V. Figueiredo, F.J. Duarte, F. Rodrigues, Z. Vale, C. Ramos, J.B. Gouveia, Electric customer characterization by clustering, Presented at the ISAP 2003, Lemnos, Greece, August, 2003.

[9] D. Gerbec, S. Gasperic, F. Gubina, Determination and allocation of typical load profiles to the eligible consumers, Presented at the IEEE Bologna Power Tech Conference, Bologna, Italy, June 23–26, 2003.

[10] D. Gerbec, S. Gasperic, I. Smon, F. Gubina, Allocation of the load profiles to consumers using probabilistic neural networks, IEEE Trans. Power Syst. 20 (2005) 548–555.

[11] D. Gerbec, S. Gasperic, I. Smon, F. Gubina, Determining the load profiles of consumers based on fuzzy logic and probability neural networks, IEE Proc. Gener. Trans. Distrib. 151 (2004) 395–400.

[12] S. Theodoridis, K. Koutroumbas, Pattern Recognition, 1st ed., Academic Press, New York, 1999.

[13] M. Kitayama, R. Matsubara, Y. Izui, Application of data mining to customer profile analysis in the power electric industry, Presented at the IEEE PES Winter Meeting 2002, New York, USA, 2002.

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

[15] M. Petrescu, M. Scutariu, Load diagram characterization by means of wavelet packet transformation, in: Proceedings of the 2nd Balkan Conference, Belgrade, Yugoslavia, June, 2002, pp. 15–19.

[16] E. Carpaneto, G. Chicco, R. Napoli, M. Scutariu, Electricity customer classification using frequency-domain load pattern data, Electr. Power Energy Syst. 28 (2006) 13–20.

[17] C.S. Chen, J.C. Hwang, C.W. Huang, Application of load survey systems to proper tariff design, IEEE Trans. Power Syst. 12 (1997) 1746–1751.

[18] D.L. Davies, D.W. Bouldin, A cluster separation measure, IEEE Trans. Pattern Anal. Mach. Intel. 2 (1979) 224–227.

[19] D. Hand, H. Manilla, P. Smyth, Principles of Data Mining, The M.I.T. Press, Cambridge, Massachusetts, London, England, 2001.

[20] S. Haykin, Neural Networks: A Comprehensive Foundation, 2nd ed., Prentice Hall, NJ, 1994.

[21] T. Kohonen, Self-Organization and Associative Memory, 2nd ed., Springer–Verlag, NY, 1989.

[22] K.F. Thang, R.K. Aggarwal, A.J. McGrail, D.G. Esp, Analysis of power transformer dissolved gas data using the self-organizing map, IEEE Trans. Power Deliver. 18 (2003) 1241–1248.

[23] SOM Toolbox for MATLAB 5, Helsinki Univ. Technology, Helsinki, Finland, 2000.