Power System Typical Load Profiles Using A New Pattern Recognition Methodology

G. J. TSEKOURAS[1,2], F.D. KANELLOS[1,2], V.T. KONTARGYRI[1,2],
E.S. KARANASIOU[1], A.D. SALIS[1], N. E. MASTORAKIS[2]
[1]School of Electrical and Computer Engineering
National Technical University of Athens
9 Heroon Polytechniou Street, Zografou, Athens
[2]Department of Computer Science,
Hellenic Naval Academy
Terma Hatzikyriakou, Piraeus, Greece
GREECE
Email: tsekouras_george_j@yahoo.gr, kanellos@mail.ntua.gr, vkont@central.ntua.gr,
ikaran@esd.ece.ntua.gr, anastasios.salis@gmail.com, mastor@wseas.org

Abstract: - In this paper a new pattern recognition methodology is described for the classification of the daily chronological load curves of power systems, in order to estimate their respective representative daily load profiles. It is based on pattern recognition methods, such as k-means, fuzzy k-means and hierarchical clustering, which are properly adapted. The parameters of each clustering method are properly selected by an optimization process using the ratio of within cluster sum of squares to between cluster variation (WCBCR) as an adequacy measure. This methodology is applied for the Greek power system, from which is proved that the separation between work days and non-work days for each season is not so enough descriptive.

Key-Words: - Load profiles, clustering algorithms, fuzzy k-means, hierarchical clustering, k-means, pattern recognition

1 Introduction

In a deregulated electricity market, load profiles categorization can be useful to the power systems and their customers. The first ones can be used for load leveling, demand side management, fluctuation smoothing and load forecasting. On the other hand the customers can participate into the competitive electricity market using demand side bidding mechanism based on powerful technologies of the energy storage systems [1] or proper tariff selection [2].

In order to carry out this classification, the chronological load curves per year, season or month can be used. During the last years, a significant research effort has been devoted to load curves classification, in order to solve the short-term load forecasting of anomalous days [3-4] and to cluster the customers of the power systems [5-10]. The clustering methods used so far are the self-organizing map [3-5], the "modified follow the leader" [5], the k-means [5], the fuzzy k-means [5-7] and the average and Ward hierarchical methods [5-7]. These methods generally belong to pattern recognition techniques [8]. Alternatively, the customers classification problem can be solved by using data mining [9], frequency-domain data [10] etc. The most commonly used respective adequacy measures are the mean index adequacy [6], the clustering dispersion indicator [5-6], the similarity matrix indicator [6], the Davies-Bouldin indicator [4-6], the modified Dunn index [5], the scatter index [5] and the mean square error [7].

The objective of this paper is to present a new methodology for the classification of the daily chronological load curves for power systems. Specifically, the respective load curves set is organized into well-defined and separate classes, in order to successfully describe the demand behavior of a power system. The proposed methodology compares the results obtained by certain clustering techniques (k-means with special weights initialization, fuzzy k-means and seven hierarchical agglomerative clustering methods) using the ratio of within cluster sum of squares to between cluster variation (WCBCR) in order to measure the adequacy. The basic aspects of this methodology are:

- the estimation of the typical days through the study period and the respective representative typical daily load profiles for the power system;
- 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 comparison of the clustering algorithms performance for the used adequacy measure.

The proposed methodology is applied in the Greek power system, although it is applicable to any power system, leading to reliable results.

2 Proposed Pattern Recognition Methodology for the Classification of Load Curves of Power System

The classification of daily chronological load curves of power system is achieved by applying the pattern recognition methodology, as shown in Fig. 1.

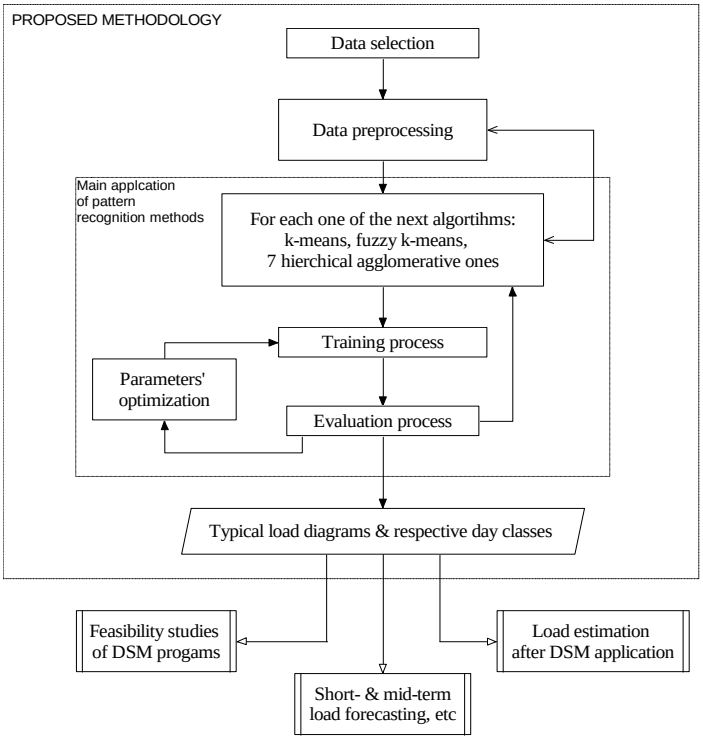

Fig. 1. Flow diagram of pattern recognition methodology for the classification of daily chronological load curves of power system

The main steps are the following:

- *Data and features selection*: The active and reactive energy values are registered (in MWh and Mvarh) for each time period in steps of 1 hour. The daily chronological load curves are determined for the study period.
- *Data preprocessing:* The load diagrams 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 are uncorrected.
- *Main application of pattern recognition methods*: For the load diagrams, a number of clustering algorithms (k-means, fuzzy k-means and hierarchical clustering) is applied. Each algorithm is trained for the set of load diagrams and evaluated according to the ratio of within cluster sum of squares to between cluster variation. The parameters of the algorithms are optimized, if necessary. The developed methodology uses the clustering methods that provide the most satisfactory results.

The results of the developed methodology can be used for power system short-term and mid-term load forecasting, energy trades, techno-economic studies of the energy efficiency and demand side management programs and the respective load estimation after the application of these programs.

3 Mathematical Modeling of Clustering Methods and Clustering Validity Assessment

3.1 General

In the study case of the chronological typical load curves of a power system a number of N analytical daily load curves is given. The main target is to determine the respective sets of days and load patterns. Generally N is defined as the population of the input vectors, which are going to be clustered. The $\vec{x}_\ell = \left(x_{\ell 1}, x_{\ell 2}, ...x_{\ell i}, ...x_{\ell d}\right)^T$ symbolizes the ℓ-th input vector and d its dimension, which equals to 24 (the load measurements are taken every hour). The corresponding set is given by $X = \{\vec{x}_\ell : \ell = 1,..., N\}$. It is worth mentioning that $x_{\ell i}$ are normalized using the higher and lower values of all elements of the original input patterns set, in order to have better results from the application of clustering methods.

Each classification process makes a partition of the initial N input vectors to M clusters. The j-th cluster has a representative, which is the respective load profile and is represented by the vector $\vec{w}_j = \left(w_{j1}, w_{j2}, ..., w_{ji}, ...w_{jd}\right)^T$ of d dimension. The vector $\vec{w}_j$ expresses the cluster center. The corresponding set is the classes set, which is defined by $W = \{\vec{w}_k, k = 1,...M\}$. 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 distance forms are defined:

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

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

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:

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

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:

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

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

3.2 K-means model

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 [3], while in the developed algorithm the w_{ji} of the *j*-th center is initialized as:

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

where *a* and *b* are properly calibrated parameters. During epoch *t* for each training vector $\vec{x}_\ell$ its Euclidean distances $d\left(\vec{x}_\ell,\vec{w}_j\right)$ 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:

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

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 variation of the weights is not significant. The algorithm's main purpose is to minimize the error function:

$$J=\frac{1}{N}\sum_{\ell=1}^{N}d^2\left(\vec{x}_\ell,\vec{w}_{k:\vec{x}_\ell\in\Omega_k}\right) \quad (6)$$

The main difference compared to 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:

$$\sum_{j=1}^{M}u_{\ell j}=1,\ u_{\ell j}:0\le u_{\ell j}\le 1,\forall j \quad (7)$$

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

$$J_{fuzzy}=\frac{1}{N}\sum_{j=1}^{M}\sum_{\ell=1}^{N}u_{\ell j}\cdot d^2\left(\vec{x}_\ell,\vec{w}_j\right) \quad (8)$$

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

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

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

where *q* is the *amount of fuzziness* in the range $(1,\infty)$ which increases as fuzziness decreases. 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 [8]. 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:

$$\begin{aligned}d\left(C_q,C_s\right)&=a_i\cdot d\left(C_i,C_s\right)+a_j\cdot d\left(C_j,C_s\right)\\&+b\cdot d\left(C_i,C_j\right)+c\cdot\left|d\left(C_i,C_s\right)-d\left(C_j,C_s\right)\right|\end{aligned} \quad (11)$$

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 (4).

The basic algorithms, which are going to be used in our case, are the *single link* algorithm (*SL*), the *complete link* algorithm (*CL*), the *unweighted pair group method average* algorithm (*UPGMA*), the *weighted pair group method average* algorithm (*WPGMA*), the *unweighted pair group method centroid* algorithm (*UPGMC*), the *weighted pair group method centroid* algorithm (*WPGMC*), the *Ward or minimum variance* algorithm (*WARD*) (for more details see [8]).

3.6 Adequacy measures

In order to evaluate the performance of the clustering algorithms and to compare them with each other the *ratio of within cluster sum of squares*

to between cluster variation (WCBCR) is used as an adequacy measure [11]. It depends on the sum of the distance square between each input vector and its cluster representative vector, as well as the similarity of the clusters centres:

$$WCBCR = \sum_{k=1}^{M} \sum_{\vec{x}_\ell \in \Omega_k} d^2\left(\vec{w}_k, \vec{x}_\ell\right) \Big/ \sum_{1 \le q < p}^{M} d^2\left(\vec{w}_p, \vec{w}_q\right) \quad (12)$$

The success of the various algorithms for the same final number of clusters is expressed by having small values of the adequacy measure. By increasing the number of clusters WCBCR decreases. It is noted that in eq. (12), M is the number of the clusters without the dead clusters (clusters for which the sets are empty).

4 Application of the Proposed Methodology

4.1 Case study

The developed methodology is applied on the Greek power system, analytically for the summer of the year 2000 and concisely for the period of years 1985-2002 per epoch and per year. The data used are hourly load values for the respective period, which is divided into two epochs: summer (from April to September) and winter (from October to March of the next year). In the case of the summer of the year 2000, the respective set of the daily chronological curves has 183 members, from which none is rejected through data pre-processing. In the next paragraphs the application of each clustering method is analyzed.

4.2 Application of the k-means

The proposed model of the *k*-means method is 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) are examined. For ten clusters the best results refer to the pair $(a,b)=(0.14,0.61)$. The alternative model is the classical one with the random choice of the input vectors during the centers' initialization. For the classical *k*-means model 100 executions are carried out and the best results for each index are registered. The superiority of the proposed model applies in all cases of clusters (see Fig. 4), while a second advantage is the convergence to the same results for the respective pairs (a,b), which can not be achieved using the classical model.

4.3 Application of the fuzzy k-means

In the fuzzy *k*-means algorithm the results of the adequacy measure depend on the amount of fuzziness increment. In fig. 2 *WCBCR* is indicatively presented for different number of clusters for three cases of $q=\{2,4,6\}$. The best results are given by $q=4$. For ten clusters the respective pair (a,b) is (0.18,0.62). It is noted that the initialization of the respective weights is similar to the proposed k-means.

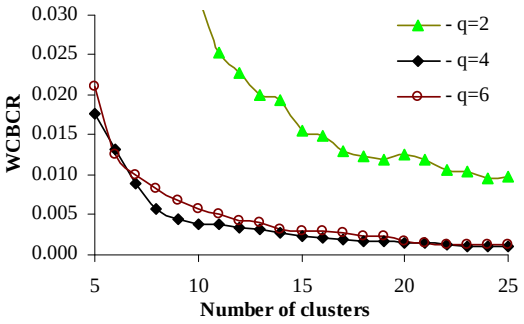

Fig. 2. *WCBCR* for the fuzzy k-means method for the set of 183 load curves of the summer of the year 2000 with q=2, 4, 6 for 5 to 25 clusters

4.3 Application of hierarchical agglomerative algorithms

In the case of the seven hierarchical models the best results are presented in fig. 3.

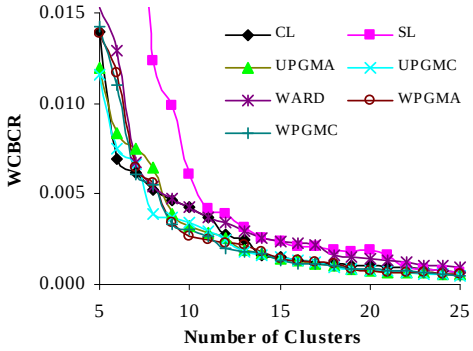

Fig. 3. *WCBCR* for the 7 hierarchical clustering algorithms for the set of 183 load curves of the summer of the year 2000 for 5 to 25 clusters

4.5 Comparison of clustering models & adequacy indicators

In fig. 4 the best results achieved by each clustering method (proposed k-means, classical k-means, fuzzy k-means and hierarchical algorithms) are depicted. The proposed k-means model has the smallest values for the *WCBCR* indicators. The improvement of the adequacy indicator is significant until 10 clusters. After this value the behavior is gradually stabilized. It can also be estimated graphically by using the rule of the "knee", which gives values between 8 to 10 clusters (see fig. 5).

Having also taken into consideration that the analogy of the computational training time for the

under study methods is 0.05:1:36 (hierarchical: proposed k-means: fuzzy k-means:), the use of the hierarchical and k-means models is proposed. It is mentioned that the computational training time for the proposed k-means method is approximately 20 minutes for a Pentium 4, 1.7 GHz, 768 MB.

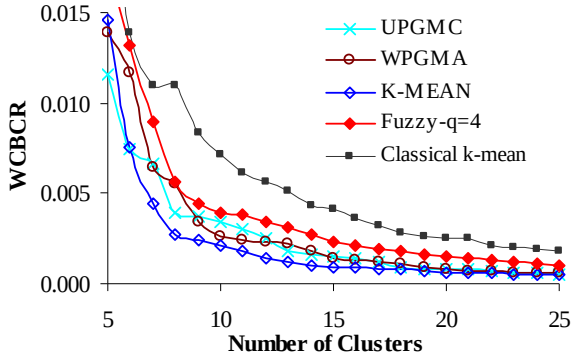

Fig. 4. The best results of each clustering method for the set of 183 load curves of the summer of the year 2000 for 5 to 25 clusters

4.7 Representative daily load curves of the summer of the year 2000 for the Greek power system

The results of the respective clustering for 10 clusters using the proposed k-means model with the optimization of the *WCBCR* indicator are presented in Table 1 and in Fig. 6 respectively.

Specifically, the cluster 1 represents Easter, the cluster 2 Holy Friday and Monday after Easter, the cluster 3 the Sundays of April, May, early June and September, Holy Saturday and Labor day. The cluster 4 contains the workdays of very low demand (during April, early May and September) with normal temperatures (22-28^{o}C) and Saturdays of April, May, early June and September, while the cluster 5 includes the workdays of low demand and Sundays of high peak load demand during the hot summer days. The cluster 6 represents the workdays of medium peak load demand and Saturdays of high peak load demand, while the clusters 7 to 10 mainly involves workdays with gradually increasing peak load demand.

TABLE 1
RESULTS OF THE PROPOSED K-MEANS MODEL WITH OPTIMIZATION TO *WCBCR* FOR 10 CLUSTERS FOR A SET OF 183 LOAD CURVES OF THE SUMMER OF THE YEAR 2000 FOR THE GREEK POWER SYSTEM

Load cluster	Day (1 for Monday, 2 for Tuesday etc.)							Days per cluster
	1	2	3	4	5	6	7	
1	0	0	0	0	0	0	1	1
2	1	0	0	0	1	0	0	2
3	0	1	0	0	0	2	13	16
4	9	8	9	8	7	12	2	55
5	4	3	2	3	4	4	8	28
6	4	6	6	4	3	7	1	31
7	4	3	4	6	6	0	1	24
8	4	3	2	3	3	2	0	17
9	0	2	3	1	2	0	0	8
10	0	0	0	1	0	0	0	1

4.8 Application of the Proposed Methodology for the Greek Power System Per Season and Per Year for the time period 1985-2002

The same process is repeated for the summers (April–September) and the winters (October-March) for years 1985-2002. The load curves of each season are qualitatively described by using 8-10 clusters. The methodology is also applied for each year during the period 1985-2002, where the load curves are qualitatively described by using 15-20 clusters. The performance of these methods is presented in Table 2 by indicating the number of seasons and of the years which achieve the best value of adequacy measure respectively. The comparison of the algorithms shows that the developed k-means method achieves a better performance for *WCBCR* measure in both cases (seasons & years).

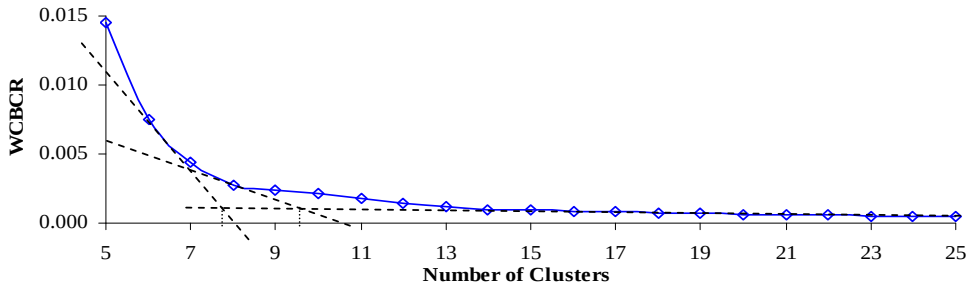

Fig. 5. Indicative estimation of the necessary clusters for the typical load daily chronological curves of the summer of the year 2000 for the *WCBCR* adequacy indicator

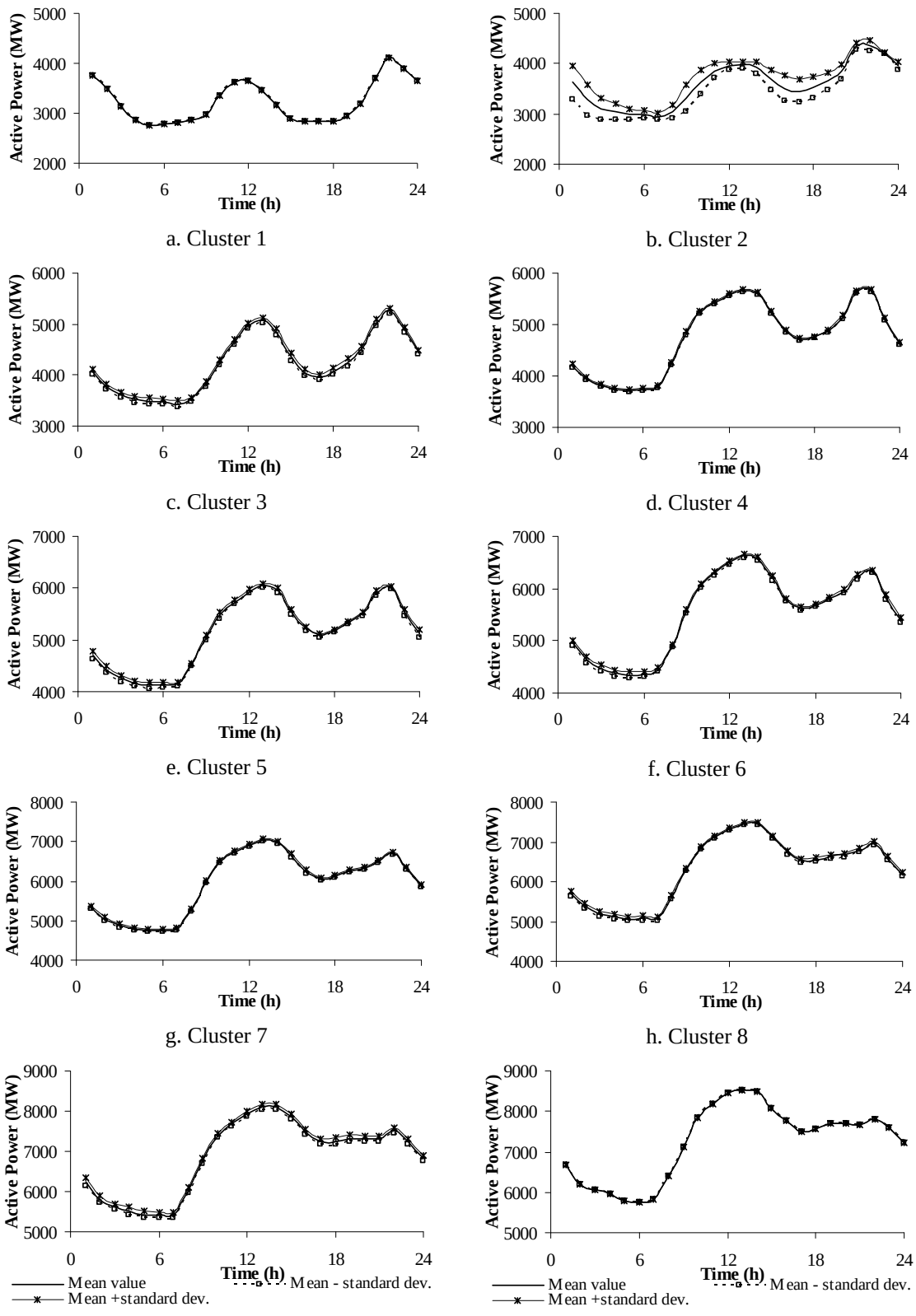

Fig. 6. Typical daily chronological load curves for the set of 183 curves of the summer of the year 2000 for the Greek power system using proposed k-means model with optimization to *WCBCR*

The main disadvantage of the load curves classification per year is that each cluster does not contain the same family of days during the time period under study. I.e. if 20 clusters are selected to represent the load demand behavior of the Greek power system per year, the 20th cluster will contain

the workdays with the highest peak load demand of the winter for the years 1985-1992 and that of summer for the rest years. In order to avoid this problem, the classification per season is proposed.

TABLE 2
COMPARISON OF THE CLUSTERING MODELS FOR THE SETS OF LOAD CURVES OF THE GREEK POWER SYSTEM PER SEASON & PER YEAR FOR THE TIME PERIOD 1985-2002

Methods	*Clustering per season*	*Clustering per year*
Proposed k-means	*30*	*14*
Classical k-means	0	0
Fuzzy k-means	1	0
CL	0	0
SL	0	0
UPGMA	1	0
UPGMC	3	2
WARD	0	0
WPGMA	0	0
WPGMC	2	2

5 Conclusions

This paper presents an efficient pattern recognition methodology for the study of the load demand behavior of power systems. The unsupervised clustering methods can be applied, such as the k-means, fuzzy k-means and hierarchical methods. The performance of these methods is evaluated by the the ratio of within cluster sum of squares to between cluster variation (WCBCR) and the representative daily load diagrams along with the respective populations per each typical day are calculated. This information is valuable for the electric companies, because it facilitates the load forecasting and the techno-economic studies of demand side management programs. From the respective application for Greek power system it is concluded that 8 to 10 clusters are necessary for the satisfactory description of the daily load curves of each season (describing the year with two seasons - winter and summer). It is practically impossible to describe the load curves satisfactory dividing the respective days into work days and non-work days, as it has been used until now.

In future the proposed methodology can be improved by used additional clustering methods, such as adaptive vector quantization (AVQ), mono-dimensional and bi-dimensional self organizing maps (SOM), and additional adequacy measures such as mean square error, mean index adequacy, clustering dispersion indicator, similarity matrix indicator, Davies-Bouldin indicator.

References:

[1] T. Sels, C. Dragu, T. Van Craenenbroeck, R. Belmans. New Energy Storage Devices for an Improved Load Managing on Distribution Level. *IEEE Power Tech Conference, 10th-13th September, Porto, Portugal,* p.6.

[2] G. J. Tsekouras, N.D. Hatziargyriou, E. N. Dialynas: "Two-Stage Pattern Recognition of Load Curves for Classification of Electricity Customers", *IEEE Transactions on Power Systems*, Vol. 22, No. 3, August 2007, pp. 1120-1128.

[3] 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.,* Vol. 11, No. 3, August 1996, pp. 1749-1756.

[4] M. Beccali, M. Cellura, V. Lo Brano, A. Marvuglia. Forecasting daily urban electric load profiles using artificial neural networks. *Energy Conversion and Management*, Vol. 45, 2004, pp. 2879-2900.

[5] G.Chicco, R. Napoli, F. Piglione. Comparisons among clustering techniques for electricity customer classification. *IEEE Trans. Power Syst.,* Vol. 21, No. 2, May 1996, pp. 933-940.

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

[8] S. Theodoridis, K. Koutroumbas, *Pattern Recognition*, 1st Edition, Academic Press, New York, 1999.

[9] V. Figueiredo, F. Rodrigues, Z. Vale, J. B. Gouveia. An electric energy consumer characterization framework based on data mining techniques. *IEEE Trans. Power Syst.,* Vol. 20, No. 2, May 2005, pp. 596-602.

[10] E. Carpaneto, G. Chicco, R. Napoli, M. Scutariou. Electricity customer classification using frequency-domain load pattern data. *Electrical Power and Energy Syst.* Vol. 28, No. 1, 2006, pp. 13-20.

[11] D. Hand, H. Manilla, P. Smyth. *Principles of data mining*, The M.I.T. Press, Cambridge, Massachusetts, London, England, 2001.