Discovering Patterns in Electricity Price Using Clustering Techniques

F. Martínez Álvarez [1], A. Troncoso [2], J. C. Riquelme [1], J. M. Riquelme [3]

[1] Departamento de Lenguajes y Sistemas Informáticos
Escuela Técnica Superior de Ingeniería Informática. Universidad de Sevilla
Phone: +0034 954 552775, e-mail: fmartinez@lsi.us.es, riquelme@lsi.us.es

[2] Área de Lenguajes y Sistemas Informáticos
Escuela Politécnica Superior. Universidad Pablo de Olavide
Phone: +0034 954 977522, e-mail: ali@upo.es

[3] Departamento de Ingeniería Eléctrica
Escuela Superior de Ingenieros. Universidad de Sevilla
Phone: +0034 954 481274 e-mail: jsantos@us.es

Abstract. Clustering is a process of grouping similar elements gathered or occurred closely together. This paper presents two clustering techniques, K-means and Fuzzy C-means, for the analysis of the electricity prices time series. Both algorithms are focused on extracting useful information from the data with the aim of model the time series behaviour and find patterns to improve the price forecasting. The main objective, thus, is to find a representation that preserves the original information and describes the shape of the time series data as accurately as possible. This research demonstrates that the application of clustering techniques is effective in order to distinguish several kinds of days. To be precise, two major groups can be distinguished thanks to the clustering: the first one that includes the working days and the second one that includes weekends and festivities. Equally remarkable is the similarity shown among days belonging to a same season.

Key words

Clustering, price forecasting, time series model.

1. Introduction

It is important to obtain an approach to optimize the bidding strategies carried out by electricity-producer companies [1]. Consequently, the development of forecasting techniques is becoming increasingly relevant in the current hectic Spanish electricity-market deregulation.

This work is focused on extracting meaningful information of the prices time series by using clustering techniques. Clustering is the basis of many classification and system modelling algorithms. The main target of clustering is to generate groupings of data from a large dataset with the intention of producing an accurate representation of the behaviour of a system.
Thus, the research is based on the application of two well-known clustering methods, K-means and fuzzy clustering [2], for finding those groups of prices which show a similar behaviour under some particular conditions such as working / non-working days or seasons. Later, this information can be used for predicting how the prices will progress throughout the next day.

Other researchers have developed techniques to forecast the prices time series. Recently, A. J. Conejo et al. [3] proposed a forecasting model using the wavelet transform and ARIMA models. Equally, R. C. García et al. [4] presented a forecasting technique based on a GARCH model. In [5] a method combining Artificial Neural Networks with fuzzy logic is proposed. In [6] an adaptive non-parametric regression approach is applied to forecast the hourly Ontario energy price. In [7] a simple model based on the Weighted Nearest Neighbours methodology is presented and its performance is compared with others recently published techniques.

However, it can be stated that the forecasting techniques for the next-day electricity prices published in the current literature do not use previous clustering techniques. Consequently, it is necessary to discover patterns in the electricity prices time series to improve the prediction models.

The final goal is to provide several predictions for the price evolution curve of the subsequent day. This information would be used in optimization models in order to help to market agents to generate their optimal bidding strategies. Thus, the first objective is to determine a previous clustering over real prices population curves in order to obtain groups of days according to the price of the electricity hour by hour.

From this division, the connection between belonging to a certain cluster and the model of prediction for each cluster could be found. Therefore, it would be chosen the cluster to which the current day it belong in order to predict the prices curve of the following day. Finally, the prediction would be generated by means of the corresponding model of this cluster.

The novel and main contribution of the paper is to apply clustering techniques to the electricity prices time series to discover similar patterns. The patterns provide useful

information to improve the forecasting techniques. The time series is the variation of the price of the electricity throughout the day. The more days considered in the dataset, the more precise will be the prediction.

The rest of the paper is organized as follows. Section 2 details the two clustering techniques applied to find patterns in time series. As selecting the number of clusters results a key process, Section 3 explains the motivation for choosing the number of clusters in both algorithms. Section 4 presents all the results obtained, as well as it compares both techniques. A description of the dataset used is also shown in this section. Finally, Section 5 expounds the conclusions achieved and the future work.

2. Methodology

Clustering is a process of grouping an unlabeled set of examples into a number of clusters such that a similar pattern is associated to every cluster, that is to say, clustering operates on a set of examples that must be partitioned according to some notion of similarity. Cluster analysis techniques have been classified into two major methods:

1) *Crisp clustering (or hard clustering)* in which the boundary between clusters is fully defined.
2) *Fuzzy clustering* in which the boundary between clusters can not be clearly defined (such is the case of many real cases).

Both approaches present a large set of algorithms, most of them designed for specific problems. In this paper two different clustering-based techniques have been used in order to identify patterns of behaviour in the prices curves: K-means, representing the crisp clustering and the Fuzzy C-means (FCM) representing the fuzzy clustering.

A. k-means algorithm

K-means is a fast method to perform clustering. The basic intuition behind K-means is the continuous reassignment of objects into different clusters so that the within-cluster distance is minimized.

It uses an iterative algorithm divided in two phases to minimize the sum of point-to-centroid distances, over all *k* clusters.

In the first phase, each iteration consists of reassigning points to their nearest cluster centroid and then it recalculates the cluster centroids.

In the second phase, points are individually reassigned if doing so reduce the sum of distances; cluster centroids are recomputed after each reassignment. Each iteration consists of one pass through all the points. Both phases are summarized in Table I, which describes the k-means in terms of its basic steps.

Table I. – Outline of the k-means algorithm

STEP	DESCRIPTION
1	Decide a value for k
2	Initialize the k cluster centres
3	Assign an example to the nearest cluster centre
4	Re-calculate the k cluster centres assuming that the memberships found in step 3 are correct
5	Exit if no example changes of cluster in the last iteration. Otherwise go to step 3.

B. Fuzzy C-means algorithm

The Fuzzy C-means clustering, where C is the number of clusters to classify, the data is a technique wherein each data belongs to a cluster to some degree specified by a membership grade. It provides a method that shows how to group data points that populate some multidimensional space into a specific number of different clusters.

The FCM algorithm focuses on minimizing the value of an objective function which calculates the weighted within-group sum of squared errors. To measure the quality of the partitioning it compares the distance from an example to the current candidate cluster centre with the distance to other candidate cluster centres. Table II shows the summarized steps followed in the algorithm.

Table II. – Outline of the Fuzzy C-means algorithm

STEP	DESCRIPTION
1	Decide a value for C
2	Initialize the cluster centre matrix, $W^{(t=0)}$
3	Initialize the membership matrix, $U^{(t=0)}$
4	Increase t by one and compute $W^{(t)}$
5	Compute $U^{(t)}$
6	If $(U^{(t)} – U^{(t-1)})$ is lower than a given error stop. Otherwise go to step 4.

While these two algorithms are typically used in the literature relative to clustering approaches in time series, they present a well-known shortcoming: the number of clusters must be specified in advance. The choice of this parameter will be justified in the subsequent section.

3. Selection of the number of clusters

The number of clusters selected is one of the most critical decisions in clustering techniques. The fact of choosing a large number of clusters does not necessarily imply have a better quality of information. On the contrary, results could be unclear and could muddle the pattern recognition up. This limitation can be mitigated by testing all values of K or C clusters within a large range. Further statistical test can, then, be used to determine which value of K or C fits better. Sections 3.A and 3.B show a methodical way to select the optimal number of clusters for both techniques.

A. Number of clusters in K-means

The *silhouette* function in Matlab provides a measure of the clusters separation. Its value varies between –1 and

+1, where +1 denotes clear cluster separation and –1 marks points with questionable cluster assignment. A successful clustering has a mean silhouette value higher than 0,6 for all clusters. However, in real time series it is almost impossible to reach this value and not having negative values in the figure is usually enough to decide how many clusters have to be chosen.

Figures 1.a, 1.b and 1.c show the plotted silhouette function for 4, 5 and 6 clusters respectively for the prices of the electricity of the year 2005. The metric used was squared Euclidean distance since cosine metrics gave worse results. For further analysis, 4 clusters have been chosen due to that only one cluster has negative values and its graphical representation provides satisfactory results.

B. Number of clusters in Fuzzy C-means

The FCM clustering algorithm is sensitive to the situation of the initialization and easy to fall into a local minimum or a saddle point when iterating. To solve this problem several other techniques have been developed that are based on global optimization methods [8]. However, in many practical applications the clustering method that is used is FCM with multiple restarts to escaping from the sensibility to initial value.

The *subclust* function in Matlab finds cluster centres and it is commonly used in order to obtain the optimum number of clusters in iterative optimization-based clustering methods such as FCM.

This function estimates the cluster centres in a set of data by using the subtractive clustering method. It assumes that each data point is a potential cluster centre and calculates a measure of the likelihood that each data point would define the cluster centre, based on the density of surrounding data points. After the execution of this algorithm, it was found that 6 is the optimum number of clusters.

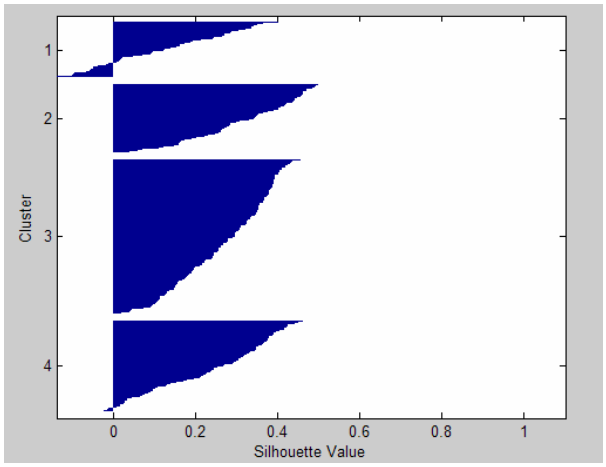

Fig. 1.a. Silhouette values with 4 clusters.

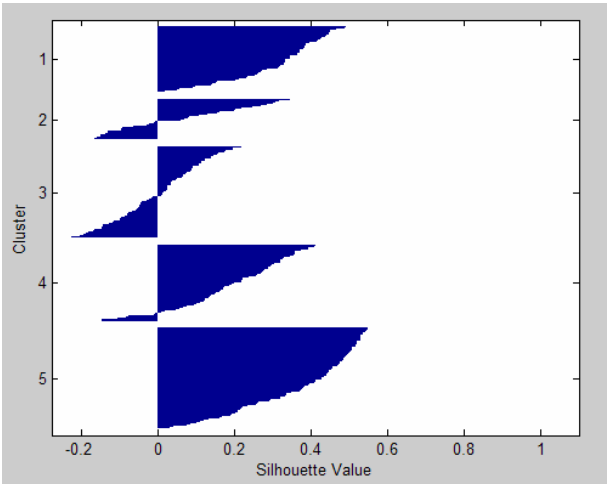

Fig. 1.b. Silhouette values with 5 clusters.

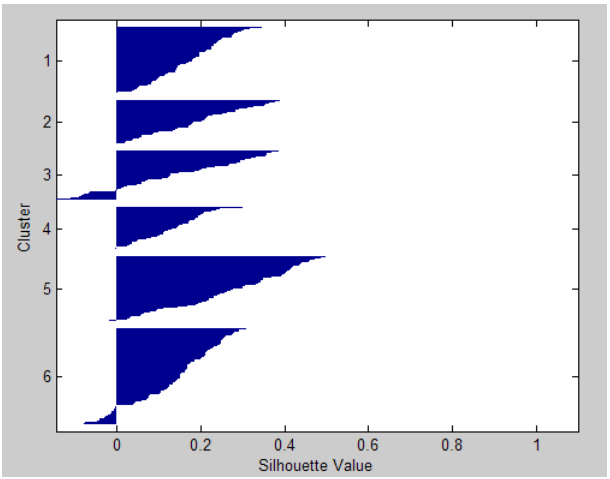

Fig. 1.c. Silhouette values with 6 clusters.

4. Results

A. Dataset description

The data source is the prices of the electricity market of mainland Spain for the year 2005 (OMEL) [9].

Before operating with the electricity prices, data normalization was carried out with the aim of avoiding the effects of the growth of the intra-annual prices. The normalization was performed by dividing the hourly prices by the average price of the whole day.

B. K-means

Figure 2 shows the year 2005 classified into 4 clusters, as justified in section 3.B, via the k-means algorithm. With just a quick look, it can be clearly differentiated two kinds of clusters: clusters 1 and 2 group all the working days and clusters 3 and 4 the weekends. Nevertheless, there are some days that have an apparently discordant behaviour. Table IV shows the percentage of days classified into the 4 clusters.

Table IV. – Grade of membership of days to clusters

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Monday	36,54%	51,92%	3,85%	7,69%
Tuesday	31,48%	57,41%	3,70%	7,41%
Wednesday	30,77%	63,46%	3,85%	1,92%
Thursday	32,69%	59,62%	5,77%	1,92%
Friday	28,85%	59,62%	3,85%	7,69%
Saturday	11,32%	0,00%	39,62%	49,06%
Sunday	0,00%	0,00%	44,23%	55,77%

There are 22 working days that have been grouped in clusters 3 or 4. A meticulous analysis reveals that most of these days were holiday. A detailed list of this fact is summarised in Table V.

Table V. – Wrong classification of working days

Nº OF DAY	DATE	FESTIVITY
6	06-01	Epiphany
70	11-03	None
75	16-03	None
77	18-03	Friday pre-Easter
82	23-03	Easter
83	24-03	
84	25-03	
87	28-03	Monday post-Easter
98	08-04	None
122	02-05	Working festivity
123	03-05	Madrid festivity
125	05-05	Long weekend 01-05
126	06-05	Long weekend 01-05
227	15-08	Assumption of Mary
231	19-08	None
235	23-08	None
285	12-10	Columbus Day
304	31-10	Long weekend 01-11
305	01-11	All Saints
340	06-12	Spanish Constitution Day
342	08-12	Immaculate Conception
360	26-12	Monday after Christmas

One comment has to be done about the first week of May. The real holiday for the Working Day is the 1st May and for the Madrid Festivity the 2nd May. However, 1st May 2005 was Sunday and both festivities were postponed one day.

With reference to weekends, there are six Saturdays that have been grouped that have been grouped as if they were working days, concretely, in cluster 1. A detailed list of these Saturdays is shown in Table VI.

Table VI. – Saturdays classified in wrong clusters

NUMBER OF DAY	DATE
169	18th June
176	25th June
183	2nd July
197	16th July
204	23rd July
211	30th July

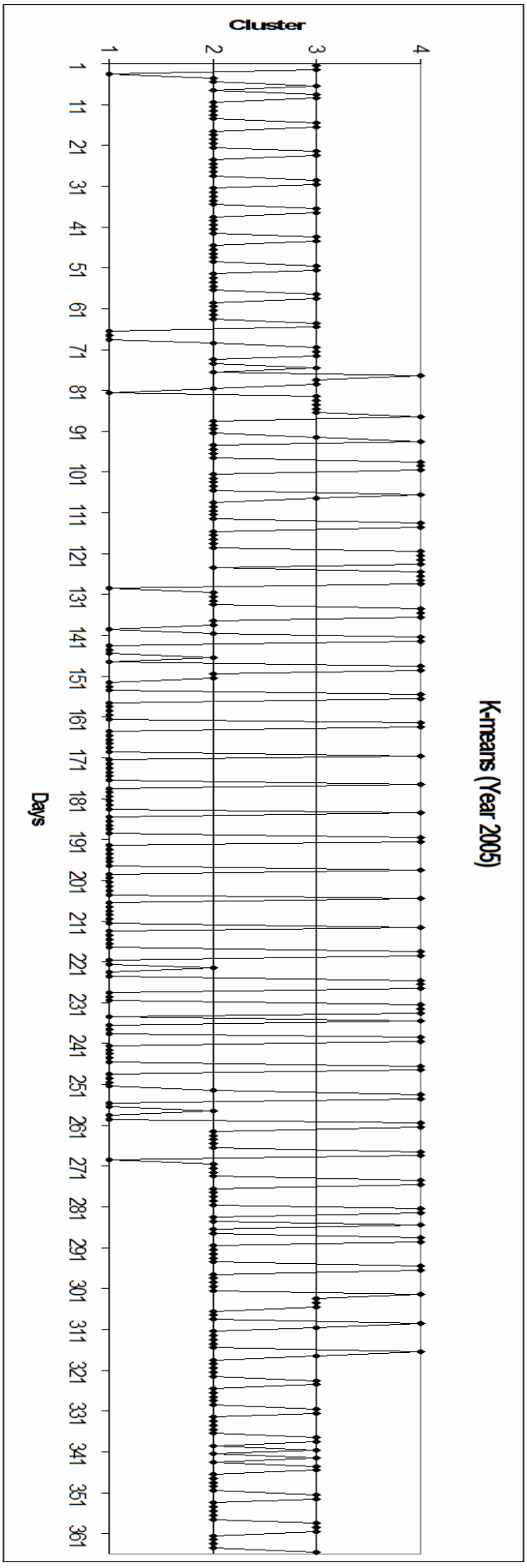

Fig. 2. Days classified into 4 clusters via K-means.

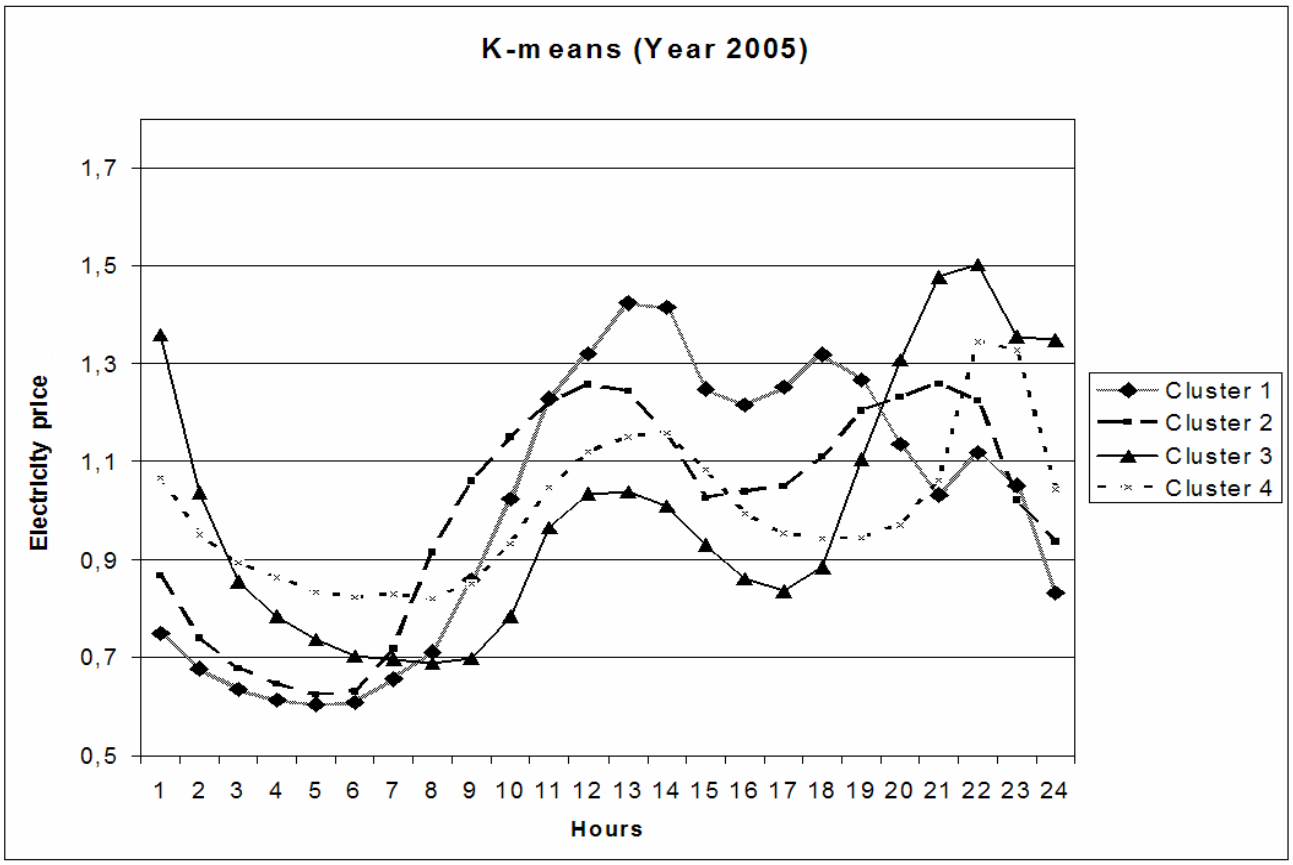

Fig. 3. Characteristic curves for clusters obtained by K-means algorithm in year 2005.

Note that almost all six Saturdays are consecutive and belong to summer, except for the 9th July that has been classified into cluster 4.

The whole year is divided into 261 working days and 104 weekends or festivities. In Table V, five days were improperly classified (11th March, 16th March, 8th April, 19th August and 23rd August). Hence, the average error in working days is 1,92% (5 days out of 261).

With regard to weekends and festivities, there are 6 Saturdays which have been improperly grouped (18th June, 25th June, 2nd July, 16th July, 23rd July and 30th July). Given that, the average error for weekends and festivities is 5,77% (6 days of out 104), the total error is 3,01% (11 days out of 365).

The following task consists in explaining when a working day belongs to cluster 1 or to cluster 2 as well as when festivities belong to cluster 3 or to cluster 4: there are three zones clearly differentiated in Figure 2 for both working days and festivities. From the 1st January until the 18th May (day number 144), most of the working days belong to cluster 2. From this day until the 20th September (day number 263) they belong to cluster 1. Finally, from the 21st September (day number 264) until the year ends the working days belong again to cluster 2.

In festivities there is a similar situation. From the 1st January until the 27th March (day number 86) most of the festivities and weekends belong to cluster 3. From this weekend until 30th October (day number 303) they belong to cluster 4. Finally, from this weekend until the year ends the festivities and weekend belong to cluster 3. Consequently, a seasonal behaviour can be observed in the energy prices time series.

The characteristic curves of each cluster are depicted by Figure 3. Especially remarkable is that curves associated to clusters 3 and 4 (weekends and festivities) have starting and ending prices higher than the ones associated to the working days (clusters 1 and 2). The first ones show their higher values in the late afternoon. It is due to people consuming more electricity all the night long during weekends. On the other hand, the second ones have their peak prices at midday when industries, commerce and enterprises are fully functioning.

C. FCM.

Figure 4 presents the six patterns found by the FCM algorithm for the energy prices of the year 2005. It can be noted that these patterns are not very different to the patterns obtained by using the K-means approach. For the representation of these curves the following methodology has been used. First, the cluster with the maximum grade of membership was assigned for every day. Then, the representation was performed like in K-means algorithm as it has depicted in Figure 4.

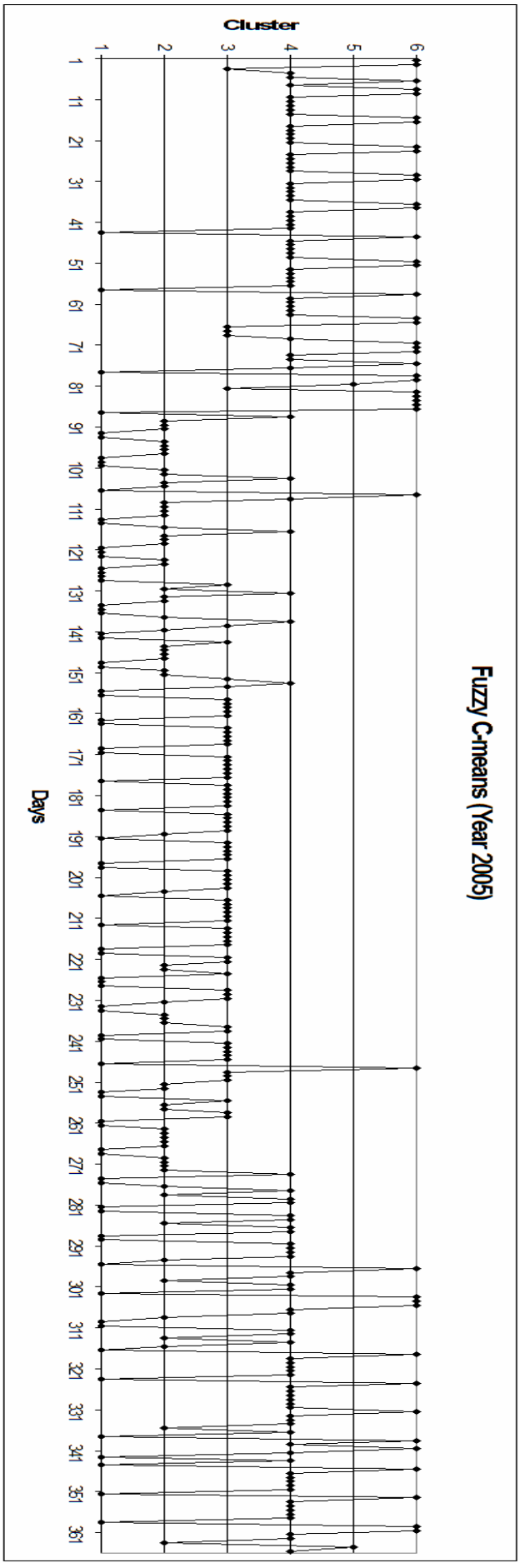

Fig. 4. Days classified into 6 clusters via FCM.

Focusing on Figure 4, it can be clearly differentiated two kinds of clusters: clusters 2, 3, 4 and 5 group all the working days while clusters 1 and 6 the weekends. Nevertheless, there are some days that have an apparently discordant behaviour. Table VII shows the percentage of days classified into the 6 clusters.

Table VII. – Grade of membership of days to clusters

	CLUSTER 1	CLUSTER 2	CLUSTER 3
Monday	7,69%	15,38%	32,69%
Tuesday	0,00%	23,08%	28,85%
Wednesday	0,00%	28,85%	26,92%
Thursday	3,85%	25,00%	26,92%
Friday	5,77%	25,00%	26,92%
Saturday	66,04%	3,77%	5,66%
Sunday	53,85%	0,00%	0,00%
	CLUSTER 4	CLUSTER 5	CLUSTER 6
Monday	38,46%	1,92%	3,85%
Tuesday	44,23%	0,00%	3,85%
Wednesday	40,38%	0,00%	3,85%
Thursday	40,38%	0,00%	3,85%
Friday	36,54%	1,92%	3,85%
Saturday	3,77%	0,00%	20,75%
Sunday	0,00%	0,00%	46,15%

There are 19 working days that have been grouped in clusters 1 or 6. Table VIII summarizes the festivities found in these days.

Table VIII. – Wrong classification of working days

Nº OF DAY	DATE	FESTIVITY
6	06-01	Epiphany
70	11-03	None
75	16-03	None
77	18-03	Friday pre-Easter
82	23-03	Easter
83	24-03	Easter
84	25-03	Easter
87	28-03	Monday post-Easter
98	08-04	None
122	02-05	Working festivity
125	05-05	Long weekend 01-05
126	06-05	Long weekend 01-05
136	16-05	None
227	15-08	Assumption of Mary
304	31-10	Long weekend 01-11
305	01-11	All Saints
340	06-12	Spanish Constitution Day
342	08-12	Immaculate Conception
360	26-12	Monday after Christmas

With reference to weekends, there are six Saturdays that have been grouped as if they were working days, concretely, in cluster 3. It is shown in Table IX.

Table IX. – Saturdays classified in wrong clusters

NUMBER OF DAY	DATE
176	25th June
183	2nd July
190	9th July
204	23rd July
211	30th July
330	26th November

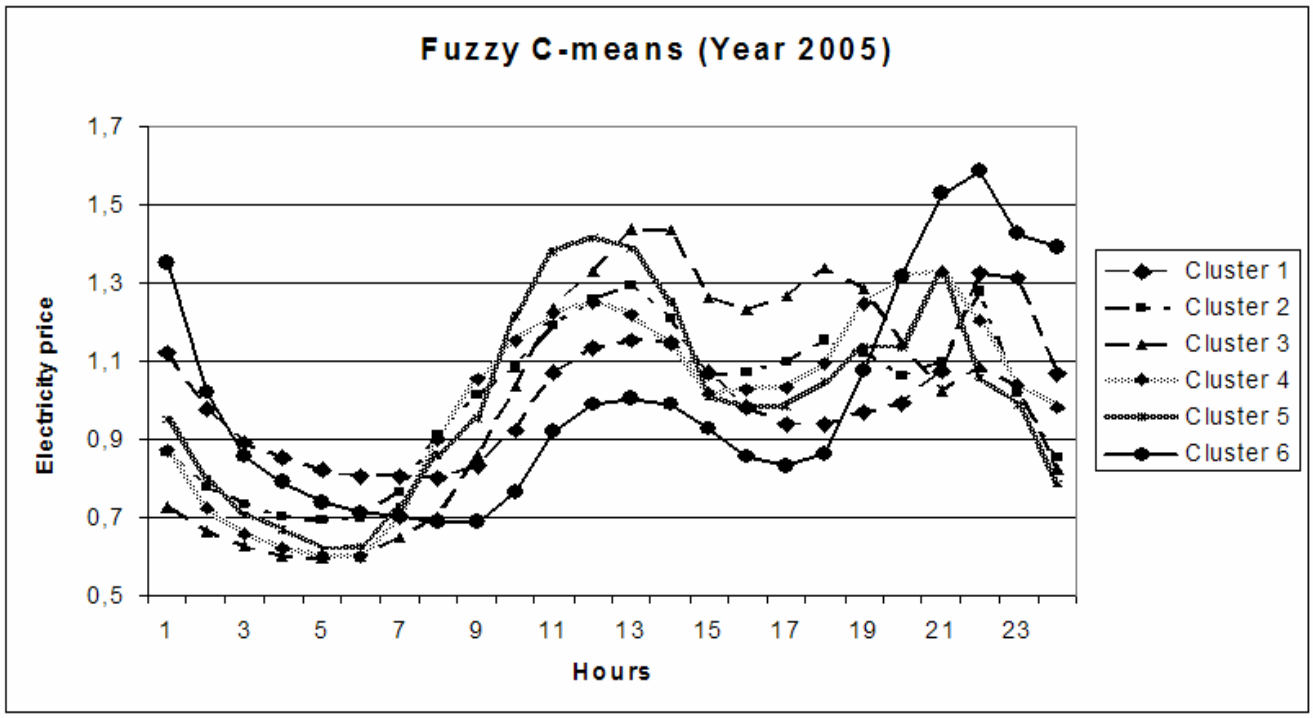

Fig. 5. Characteristic curves for clusters obtained by K-means algorithm in year 2005.

Note that almost all six Saturdays are consecutive and belong to summer, except for the 16th July that has been classified into cluster 1. The distance from 16th July to cluster 1 is 0,7910 (the cluster to which it belongs) while to cluster 2 is 0,7970 (the cluster to which it should belong, assuming that all the Saturdays in summer behave as if they were a working day).

The whole year is divided into 261 working days and 104 weekends or festivities. In table VIII, four days were improperly classified (11th March, 16th March, 8th April, 16th May). Hence, the average error in working days is 1,53% (4 days out of 261).

With regard to weekends and festivities, there are six Saturdays which have been improperly grouped (25th June, 2nd July, 9th July, 23rd July, 30th July and 26th November). There is also a festivity, Columbus Day, which has been grouped in cluster 2. Given that, the average error for weekends and festivities is 6,73% (7 days of out 104), the total error is 3,01% (11 days out of 365).

In contrast to what it happened with K-means clustering, it is not obvious to determine clear periods of the year for days to belong to a specific cluster.

The characteristic curves of each cluster are depicted by Figure 5. Especially remarkable is that curves associated to clusters 1 and 6 (weekends and festivities) have starting and ending prices higher than the ones associated to the working days (clusters 2, 3, 4 and 5). The first ones show their higher values in the late afternoon. It is due to people consuming more electricity all the night long during weekends. On the other hand, the second ones have their peak prices at midday when industries, commerce and enterprises are fully functioning.

5. Conclusions

It has been proven that utilising clustering techniques in the prices time series is as powerful as useful. Two algorithms have been used in order to classify the electricity price curves of the Spanish Market: K-means and Fuzzy C-means. The cluster analysis carried out via both K-means and Fuzzy C-means algorithms yielded very relevant information: working days have behaviour diametrically opposite to weekend and festivities. The average error committed in their classification was 3,01%. Only 11 days of the year 2005 were improperly grouped which means a great degree of accuracy of both techniques with either 4 (K-means) or 6 (Fuzzy C-means) clusters.

The results obtained may be extremely profitable. Future works will be directed in the prediction of day-ahead prices once known the previous clustering. In short, N clusters will be obtained and different models will be applied to every cluster to improve the quality of the market price forecasting.

Acknowledgements

The authors would like to acknowledge the financial support from the Spanish Ministry of Science and Technology, projects TIN2004-00159 and ENE-2004-03342/CON, and from the Junta de Andalucía, project P05-TIC-00531.

References

[1] M. A. Plazas, A. J. Conejo and F. J. Prieto, "Multimarket Optimal Bidding for a Power Producer", IEEE Transactions on Power Systems, vol. 20, no. 4, November 2005.
[2] Rui Xu and Donald C. Wunsch II, “Survey of Clustering Algorithms”, IEEE Transactions on Neural Networks, vol. 16, no. 3, May 2005.
[3] A. J. Conejo, M. A. Plazas, R. Espínola and A. B. Molina, “Day-Ahead Electricity Price Forecasting Using the Wavelet Transform and ARIMA Models”, IEEE Transactions on Power Systems, vol. 20, no. 1, May 2005.
[4] R. C. García, J. Contreras, M. van Akkeren and J. B. C. García, “A GARCH Forecasting Model to Predict Day-Ahead Electricity Prices”, IEEE Transactions on Power Systems, vol. 20, no. 1, May 2005.
[5] N. Amjady, “Day-Ahead Price Forecasting of Electricity Markets by a New Fuzzy Neural Network”, IEEE Transactions on Power Systems, vol. 21, no. 2, May 2006.
[6] H. Zeripour, K. Bhattacharya and C. A. Cañizares, “Forecasting the Hourly Ontario Energy Price by Multivariate Adaptive Regression Splines”, IEEE Power Engineering Society General Meeting, 2006.
[7] A. Troncoso, J. M. Riquelme, A. Gómez Expósito, J. L. Martínez Ramos and J. C. Riquelme, “Electricity Market Price Forecasting Based on Weighted Nearest Neighbors Techniques”, IEEE Transactions on Power Systems, in press, 2006.
[8] W. Wang, Y. Zhang, Y. Li and Xiaona Zhang, “The Global Fuzzy C-Means Clustering Algorithm”, Proceedings of the 6th World Congress on Intelligent Control and Automation, June 2006.
[9] OMEL. Market Operator of the Electricity Market of Mainland Spain. [Online] Available: http://www.omel.es.