Partitioning-Clustering Techniques Applied to the Electricity Price Time Series

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

[1]Area of Computer Science. Pablo de Olavide University, Spain
{fmaralv,ali}@upo.es
[2]Department of Computer Science. University of Seville, Spain
riquelme@lsi.us.es
[3]Department of Electrical Engineering. University of Seville, Spain
jsantos@us.es

Abstract. Clustering is used to generate groupings of data from a large dataset, with the intention of representing the behavior of a system as accurately as possible. In this sense, clustering is applied in this work to extract useful information from the electricity price time series. To be precise, two clustering techniques, K-means and Expectation Maximization, have been utilized for the analysis of the prices curve, demonstrating that the application of these techniques is effective so to split the whole year into different groups of days, according to their prices conduct. Later, this information will be used to predict the price in the short time period. The prices exhibited a remarkable resemblance among days embedded in a same season and can be split into two major kind of clusters: working days and festivities.

Key words: Clustering, electricity price forecasting, time series, day-ahead energy market

1 Introduction

Due to the Spanish electricity-market deregulation, a will of obtaining optimized bidding strategies has recently arisen in the electricity-producer companies [13]. In that way, forecasting techniques are acquiring significant importance. Thus, this research lies in extracting useful information of the prices time series by using clustering techniques. In this work two well-known clustering techniques [15], K-means and Expectation Maximization (EM), are applied to prices time series in order to find those days which show a similar behavior. These labeled days will be used to forecast the day-ahead price in future work.

Several forecasting techniques have already been used in forecasting miscellaneous electricity time series recently. Indeed, A. J. Conejo et al. [2] used the wavelet transform and ARIMA models and R. C. García et al. [4] presented a forecasting technique based on a GARCH model for this purpose. A mixing of Artificial Neural Networks and fuzzy logic were proposed in [1], while an adaptive non-parametric regression approach is handled in [17]. A model based on

the Weighted Nearest Neighbors methodology is presented in [14]. With the aim of dealing with the spike prices, [6] proposed a data mining approach based on both support-vector machine and probability classifier. In [5] mixed models were proposed to obtain the appropriate length of time to use for forecasting prices.

However, none of them used clustering techniques applied to prices time series as a previous stage. The novel and main contribution of this paper is to apply clustering to the electricity prices time series in order to discover behavior's patterns, as a first step to improve forecasting techniques. Therefore, this work tackle the problem in a framework based on non-supervised learning techniques, which will enhance the prices prediction accuracy. The input data is the hourly variation of the price of the electricity throughout the day and is available on [12].

The rest of the paper is organized as follows. In Section 2 the algorithms used, K-means and EM, are described. It is also discussed the number of clusters selected for the analysis. Section 3 shows the results obtained by each method, giving a measure of the quality of them. Finally, Section 4 expounds the conclusions achieved and gives the clues for future work.

2 Partitioning-Clustering Techniques

It has been already demonstrated that partitioning-clustering techniques perform better classifications than fuzzy clustering when electricity prices are considered [11]. In this section two methods are presented, K-means and EM, in order to choose the best algorithm among the partitioning ones. The number of clusters to be generated is one of the most critical parameters, insofar as a too high number could turn the results unclear and muddle the pattern recognition up. Consequently, this optimal number will be widely discussed for each algorithm.

2.1 K-means clustering technique

K-means [10] 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. The procedure can be summarized as follows:

1. *Phase 1.* In each iteration (evaluation of all the points) every point is reassigned to their closest cluster center. Then the clusters centers are recalculated.
2. *Phase 2.* Points are reassigned only if the sum of distances is reduced. The clusters centers are recalculated after each reassignment.

Selecting the number of clusters. The *silhouette function* [7] provides a measure of the quality of the clusters' separation obtained by using the K-means

algorithm. In an object i belonging to cluster C_k, the average dissimilarity of i to all other objects of C_k is denoted by $c_k(i)$. Analogously, in cluster C_m, the average dissimilarity of i to all objects of C_m is called $dis(i, C_m)$. After computing $dis(i, C_m)$ for all clusters $C_m \neq C_k$, the smallest one is selected: $c_m(i) = min\{dis(i, C_m)\}$, $C_m \neq C_k$. This value represents the dissimilarity of i to its neighbor cluster. Thus, the silhouette $silh(i)$ is given by the following equation:

$$silh(i) = \frac{c_k(i) - c_m(i)}{max\{c_k(i), c_m(i)\}} \tag{1}$$

The $silh(i)$ can vary between -1 and $+1$, where $+1$ denotes clear cluster separation and -1 marks points with questionable cluster assignment. If cluster C_k is a singleton, then $silh(i)$ is not defined and the most neutral choice is to set $silh(i) = 0$. The objective function is the average of $silh(i)$ over the N objects to be classified, and the best clustering is reached when the above mentioned function is maximized.

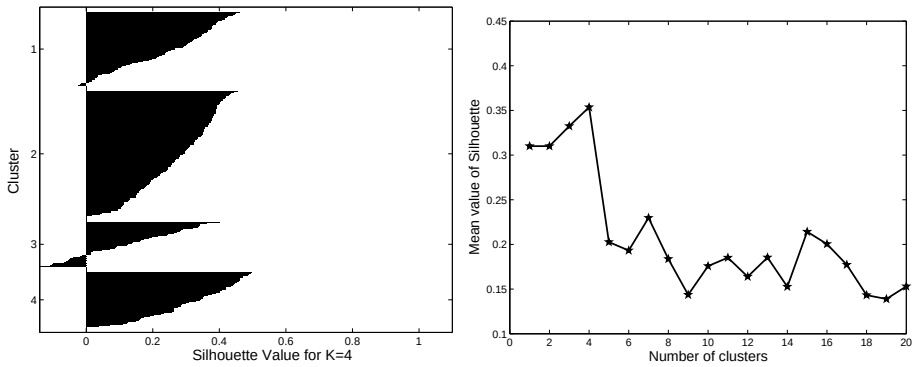

Fig. 1. Silhouette function when $K = 4$. Clusters 2 and 4 are perfectly separated as no negative values were obtained, while clusters 1 and 3 present some uncertainty. The right picture shows the mean value of *silhouette* when varying K.

The metric used to determine the *silhouette function,* shown in Figure 1, was the squared Euclidean distance since cosine metrics gave worse results. The maximum mean silhouette value obtained was 0.35, when evaluating the number of clusters from 1 to 20, and it was reached when four clusters were taken into consideration. For this reason [7], the number of clusters selected for further analysis is four ($K = 4$).

2.2 Expectation Maximization

The EM algorithm, proposed by Lauritzen in 1995 [9], is a variation the K-means. The main novelty of this technique is to obtain the previously unknown *Probability Distribution Function* (PDF) [16] of the complete dataset.

This PDF can be approximated as a linear combination of NC components, defined from certain parameters $\Theta = \cup\Theta_j$, $\forall_j = 1...NC$ that have to be found.

$$P(x) = \sum_{j=1}^{NC} \pi_j p(x; \Theta_j) \tag{2}$$

$$\sum_{j=1}^{NC} \pi_j = 1 \tag{3}$$

where π_j are the *a priori* probability of each cluster, $P(x)$ denotes the arbitrary PDF and $p(x; \Theta_j)$ the PDF of each j component. Each cluster corresponds to their respective data samples, which belong to a every single density that are combined. PDF of arbitrary shapes can be estimated by using t-Student, Bernouilli, Poisson, normal or log-normal functions. In this research, the normal distribution has been used as shape of the PDF.

The adjustment of the parameters of the model requires some fitting measure, that is to say, how well fit the data into the distribution. This measure is called data *likelihood*. Therefore, the Θ parameters have to be estimated by maximizing the *likelihood* (ML-Maximum Likelihood criterion) [3]. But what it is usually used is the logarithm of the *likelihood* (*log-likelihood*) because of its easiness to be analytically calculated. The formula of the *log-likelihood* is:

$$L(\Theta, \pi) = log\Pi_{n=1}^{NI} P(x_n) \tag{4}$$

where NI is the number of instances, which are considered to be independent one to another. The EM algorithm, thus, can be summarized in two steps:

1. *Expectation.* It uses the initial values or the ones provided by the previous iteration of the Maximization step in order to obtain different shapes (K-means only finds hyper-spherical clusters) of the desired PDF.
2. *Maximization.* It obtains new parameters values from the data provided in the previous step, maximizing the likelihood measure by using the ML method.

After few iterations, the EM algorithm tends to a local maximum of the L function. Finally, a set of clusters, defined by the parameters of the normal distribution, will be obtained.

Selecting the number of clusters. In the EM algorithm the optimum number of clusters has been obtained with *cross-validation* [8]. The cross-validation method consists in dividing the sample dataset into subsets. The analysis is performed on only one subset while the rest of subsets are used in subsequent confirmation and validation of the initial analysis.

In this research, *V-fold cross-validation* has been used and the original dataset is partitioned into ten subsets or folds ($V = 10$). Only one of these ten subsets is retained as validation data for checking the model, while the remaining nine

subsets are utilized as training data. The *cross-validation* process is performed ten times, that is to say, each of the ten subsets are used once as validation data. Finally, the ten results obtained from the folds are averaged and combined to produce a unified estimation. Figure 2 shows the evolution of the logarithm of the likelihood function (log-ML). Thus, the number of clusters selected is eleven since its log-ML value is maximum.

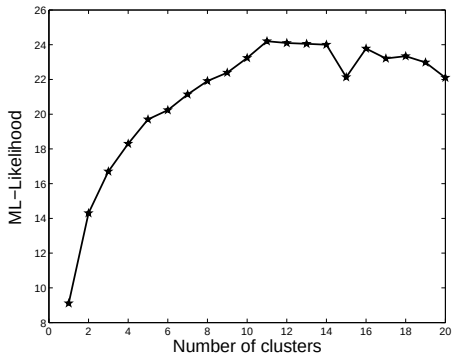

Fig. 2. Justification for the election of the number of clusters with EM.

3 Results

The K-means and EM algorithms described in the previous section have been applied in several experiments in order to obtain the forecast of the Spanish electricity price time series for the year 2005 [12].

3.1 K-means results

Figure 3 (the left one) shows the year 2005 classified into the 4 clusters. In the x axis are listed the days of the year and in the y axis the cluster to which they belong.

From this automatic classification, two kinds of clusters are easily differentiated. Working days belong to clusters 1 and 2 since they do not contain any Saturday or Sunday. Therefore, the weekends and festivities belong to clusters 3 and 4. This differentiation has been done on the basis of the following criterium. Focusing on samples 10 to 16, it can be appreciated that the 5 first samples (Monday to Friday) belong to cluster 2. On the contrary, samples 15 and 16, Saturday and Sunday respectively, belong to cluster 3 (festivities behave like weekends). This pattern is repeated all the year long but for some samples, whose membership has to be analyzed in detail.

The percentage of membership to the clusters is shown in Table 1.

Table 1. Distribution of the days in the four clusters created with K-means.

Day	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%

Table 2. Working days misclassified with K-means

N^{o} 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 Day
123	03-05	Madrid Festivity
125	05-05	Long weekend 1^{st} May
126	06-05	Long weekend 1^{st} May
227	15-08	Assumption of Mary
231	19-08	None
235	23-08	None
285	12-10	Columbus Day
304	31-10	1^{st} November long weekend
305	01-11	All Saints'
340	06-12	Spanish Constitution Day
342	08-12	Immaculate Conception
360	26-12	Monday after Christmas

Table 3. Weekends misclassified with K-means

Number of day	Date
169	18^{th} June
176	25^{th} June
183	2^{nd} July
197	16^{th} July
204	23^{rd} July
211	30^{th} July

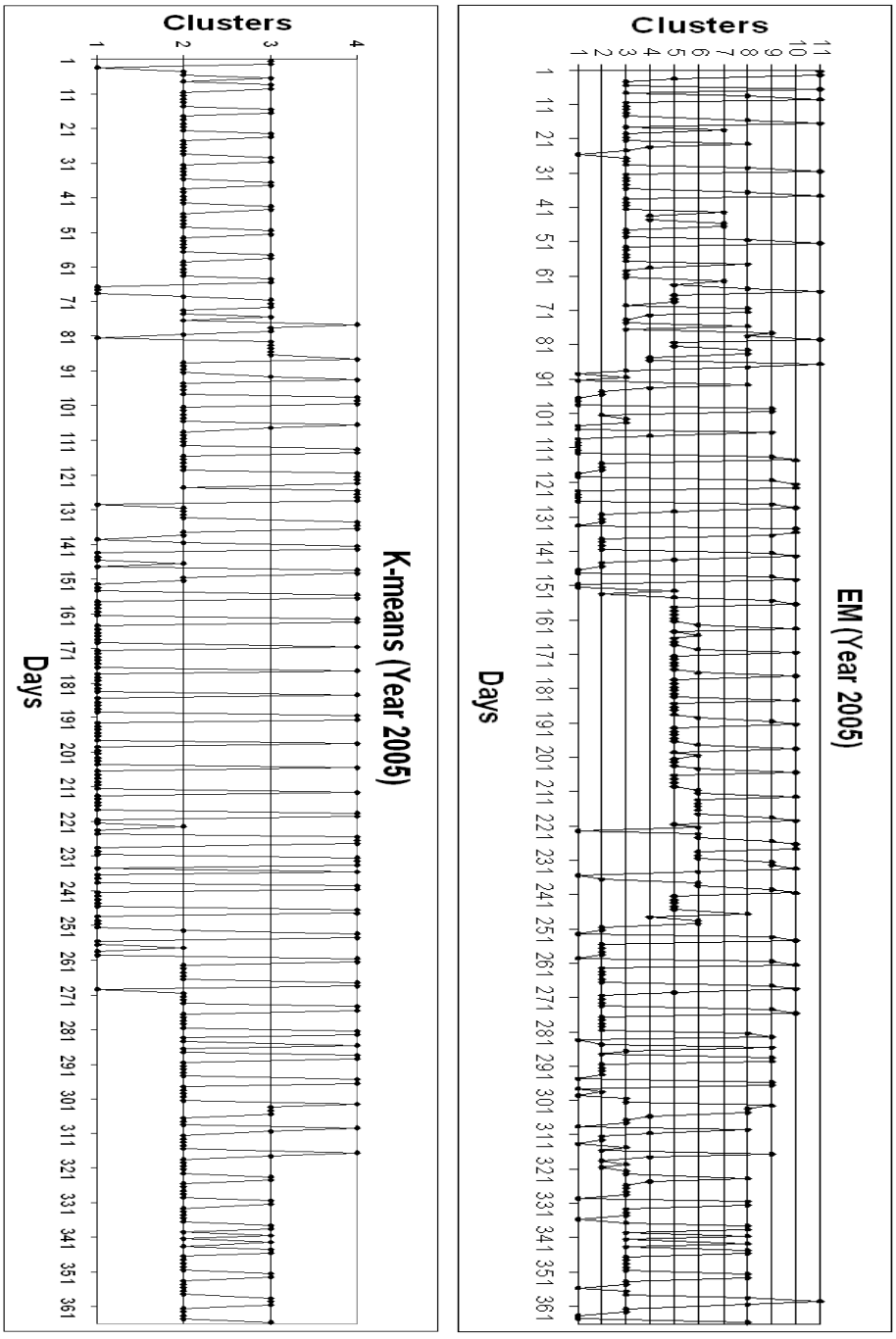

Fig. 3. Distribution of the days belonging to 2005 into the different clusters. The left figure represents the assignation with K-means and the right one with EM.

Although some days seem not to belong to the right cluster, a thorough analysis explains this phenomenon. For example, the 6^{th} day of the year was a Thursday and, according to the previous classification, it should belong to clusters 1 or 2. However, 6^{th} January is a festivity (Epiphany), therefore it behaves as if it was a weekend. For this reason it belongs to cluster 3. This situation is repeated 22 times, that is to say, there are twenty two *working days* that have been grouped in clusters 3 or 4, the clusters associated to weekends and festivities. These days are listed in Table 2.

With regard to weekends, there are six Saturdays that have been grouped as if they were working days, that is to say, they have been classified in cluster 1 (one of the clusters identified to belong to the working days) instead of being in either cluster 3 or 4, as it should belong according to the previous classification. These days are listed in Table 3.

The whole year is divided into 261 working days and 104 weekends or festivities. Only five days were not correctly classified (11^{th} March, 16^{th} March, $8t^{th}$ April, 19^{th} August and 23^{rd} August), hence, the average error in working days is 1.92% (5 days out of 261). On the other hand, there were 6 Saturdays improperly grouped. Consequently, the average error for weekends and festivities is 5.77% (6 days of out 104). Thus, the total error is 3.01% (11 days out of 365).

In Figure 3 (the left one) there are three zones clearly differentiated for both working days and festivities. From the 1^{st} January until the 18^{th} May (day number 144), most of the working days belong to cluster 2. From this day until the 20^{th} September (day number 263) they belong to cluster 1. Finally, from the 21^{st} 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 27^{th} March (day number 86) most of the festivities and weekends belong to cluster 3. From this weekend until 30^{th} 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 behavior can be observed in the energy prices time series.

3.2 EM results

Figure 3 (the right one) shows the year 2005 classified into eleven clusters via the EM algorithm. In the x axis are enumerated the days of the year and in the y axis the cluster to which they belong.

From Table 4, it can be stated that the clusters 1, 2, 3, 5 and 7 group clearly the working days since they do not contain any Saturday or Sunday. The clusters 4, 6, 8, 9, 10 and 11 the weekends and festivities, as they hardly contain Mondays, Tuesdays, Wednesdays, Thursdays or Fridays. Further division can be done is this second group. The clusters 4, 10 and 11 are mainly Sundays, while the clusters 8 and 9 are mainly Saturdays. However, the association of days to clusters with the EM algorithm is not as easy as it resulted with only four clusters. Thus, the dispersion through the clusters is higher. This fact is manifested by a higher error rate since one Saturday and sixteen working day were improperly classified or, equivalently, a 4.38% error rate was committed.

Table 4. Distribution of the days in the eleven clusters created with EM.

Cluster	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
Cluster 1	7.69%	9.62%	15.38%	15.38%	26.92%	0.00%	0.00%
Cluster 2	17.31%	25.00%	23.08%	17.31%	11.54%	0.00%	0.00%
Cluster 3	25.00%	28.85%	30.77%	34.62%	25.00%	0.00%	0.00%
Cluster 4	0,00%	1.92%	0.00%	0.00%	1.92%	3.77%	19.23%
Cluster 5	30.77%	17.31%	21.15%	17.31%	17.31%	1.89%	0.00%
Cluster 6	5.77%	11.54%	3.85%	7.69%	9.62%	11.32%	0.00%
Cluster 7	1.92%	3.85%	0.00%	1.92%	1.92%	0.00%	0.00%
Cluster 8	5.77%	1.92%	3.85%	3.85%	1.92%	39.62%	9.62%
Cluster 9	1.92%	0.00%	1.92%	0.00%	3.85%	39.62%	7.69%
Cluster 10	3.85%	0.00%	0.00%	0.00%	0.00%	1.89%	44.23%
Cluster 11	0.00%	0.00%	0.00%	1.92%	0.00%	1.89%	19.23%

In contrast to what happened in K-means, these sixteen working days do not correspond to weekends or festivities. On the contrary, this misclassification appears randomly and there are no apparent causes. Nevertheless, the Saturday wrong classified (classified into cluster 5) is, like it happened with K-means (see Table 3), the 2^{nd} July: the starting day of holidays for many Spanish people.

4 Conclusions

Partitioning-clustering techniques have been proven to be useful to find patterns in electricity price curves. The analysis carried out via both K-means and Expectation Maximization algorithms yielded relevant information insofar as they found patterns in price time series' behavior.

The average error committed in their classification was 3.01% (11 days) with K-means and 4.38% (16 days) with EM, which means a great degree of accuracy. K-means has been confirmed to be the algorithm more suitable for daily prices classification. Several factors that affect the prediction by increasing the error rate has been identified, such as the time of the day, the day of the week and the month of the year.

Future work is directed to the prediction of day-ahead prices once known the previous clustering. Therefore, the prices prediction will be handled by means of the information gathered from this clustering and used as a temporal indicator of the time series behavior. The K-means algorithm is used, thus, as a step prior to forecasting. Eventually, a label-based algorithm will be proposed with the aim of taking advantage of this extracted knowledge.

Acknowledgments. The authors want 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. N. Amjady. Day-ahead price forecasting of electricity markets by a new fuzzy neural network. *IEEE Transactions on Power Systems*, 21(2):887–896, 2006.
2. A. J. Conejo, M. A. Plazas, R. Espínola, and B. Molina. Day-ahead electricity price forecasting using the wavelet transform and ARIMA models. *IEEE Transactions on Power Systems*, 20(2):1035–1042, 2005.
3. H. Cramér. *Mathematical methods of statistics*. Princeton Univ. Press, 1946.
4. R. C. García, J. Contreras, M. van Akkeren, and J. B. García. A GARCH forecasting model to predict day-ahead electricity prices. *IEEE Transactions on Power Systems*, 20(2):867–874, 2005.
5. C. García-Martos, J. Rodríguez, and M. J. Sánchez. Mixed models for short-run forecasting of electricity prices: Application for the spanish market. *IEEE Transactions on Power Systems*, 22(2):544–552, 2007.
6. S. Guha, R. Rastogi, and K. Shim. A framework for electricity price spike analysis with advanced data mining methods. *IEEE Transactions on Power Systems*, 22(1):376–385, 2007.
7. L. Kaufman and P. J. Rousseeuw. *Finding groups in Data: an Introduction to Cluster Analysis*. Wiley, 1990.
8. R. Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of International Joint Conference on Artificial Intelligence*, pages 1137–1143, 1995.
9. S. L. Lauritzen. The EM algorithm for graphical association models with missing data. *Computational Statistics and Data Analysis*, 19(2):191–201, 1995.
10. J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the 5th Berkeley Symposium on Mathematical Statististics and Probability*, pages 281–297, 1968.
11. F. Martínez-Álvarez, A. Troncoso, J. C. Riquelme, and J. M. Riquelme. Discovering patterns in electricity prices using clustering techniques. In *Proceedings of the International Conference on Renewable Energies and Power Quality*, 2007.
12. Spanish Electricity Price Market Operator. On-line. *http://www.omel.es*.
13. M. A. Plazas, A. J. Conejo, and F. J. Prieto. Multimarket optimal bidding for a power producer. *IEEE Transactions on Power Systems*, 20(4):2041–2050, 2005.
14. A. Troncoso, J. C. Riquelme, J. M. Riquelme, J. L. Martínez, and A. Gómez. Electricity market price forecasting based on weighted nearest neighbours techniques. *IEEE Transactions on Power Systems*, 22(3):1294–1301, 2007.
15. R. Xu and D. C. Wunsch II. Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, 16(3):645–678, 2005.
16. S. Zacks. *The theory of statistical inference*. Wiley, 1946.
17. H. Zareipour, K. Bhattacharya, and C. A. Cañizares. Forecasting the hourly Ontario energy price by multivariate adaptive regression splines. *IEEE Transactions on Power Systems*, 20(2):1035–1042, 2006.