Characterization and Classification of Daily Electricity Consumption Profiles: Shape Factors and k-Means Clustering Technique

Milton Mora-Alvarez[1], *Pedro* Contreras-Ortiz[1], *Xavier* Serrano-Guerrero[1], and *Guillermo* Escrivá-Escriva[2]

[1]Grupo de Investigación en Energías, Universidad Politécnica Salesiana, Cuenca, Ecuador
[2]Institute for Energy Engineering, Universitat Politècnica de València, Valencia, Spain

Abstract. This paper exposes a method to classify the electric consumption profiles of different types of consumers, based on patterns given. The direct characteristics method is used in this paper, this method is also known as shape factors deduction (SFs) to easily define consumption profiles by using the load patterns resulting from measurements in the time domain, considering weekdays and time ranges. After the characterization of load profiles, k-means clustering technique is applied to SFs. The SFs are segmented in such a way that, in each group similar SFs are gathered. The characterization and classification of electric profiles has important applications, such as the application of specific tariffs according the consumer type, determination of optimal location of generation resources in electrical distribution systems, detection of anomalies in transmission and distribution of electricity or classify geographical areas according to electricity consumption and perform an optimum balance of feeders in electrical substations.

1 Introduction

Due to the liberalization of the electricity market, electricity trading companies have more freedom to set tariffs, by following technical and economic legal regulations. The electricity companies are developed in a profit-oriented context and they are interested in developing market strategies for the tariffs formulation. Detailed knowledge of customer consumption is essential to design specific tariff options, in which tariffs are aligned to the effective electricity use [1], [2].

The generation, transmission, distribution and commercialization companies currently have an extensive measurements database [3]. The analysis of these data allows consumer characterization, however, this task becomes complicated when the appropriate tools are not available [4].

The characterization of consumers is used for the integral planning of the system, considering alternative electricity demand management (EDM). In EDM, the effectiveness of each strategy must be evaluated by characterizing the load profile and identifying the characteristics of the demand in each type of consumer [1]. If the electric companies know in detail the demand behavior, they can improve their tariffs offer [5].

The clustering of consumers based on their activity information or commercial codes seems inefficient, since the load patterns show important differences. The classification based on the consumption pattern similarity provides more effective results [6]. The procedure to classify consumers has different stages, i) define the information that can be collected in the field, ii) choose the characteristics that are used to execute the clustering techniques, iii) use clustering techniques, iv) evaluate the effectiveness of clustering by calculating validity indicators and v) establish the final load profiles that represent a limited number of final consumers types. In order to establish a consumer clustering based on similarity aspects, there are many techniques [6].

The importance of the characterization of consumers is highlighted in [7]. In this review, an extensive review of concepts and methods for the classification in competitive electricity markets is made. In addition, a general description of classification techniques, including classical approaches, is presented. Finally, this paper is focused on developed methods that describe the electrical behavior of customers based on real cases.

In [1], an integral approach is proposed to investigate the market's strategies development based on consumers' clustering into well-defined and non-overlapping classes.

Aiming to make the classification, two sets of indexes are considered, the first set is the priority indexes, based on contractual and historical data, which are stored in the public services database, and the second set is the measured indexes, obtained through field measurements.

For the management of the historical data of the different electrical consumption patterns (ECPs). We choose unsupervised procedures or also called clusters to find characteristics to be used in the categorization. This allows obtaining an idea of the nature or structure of the data [8].

All the clustering techniques are based on mathematical criteria that assess the quality of clustering models that comprise prototypes and data partitioning. The mathematical criteria serve as cost functions that must be minimized to obtain optimal cluster solutions [9]. Such cost functions define an optimization criterion of the clustering techniques.

These techniques can be used to simplify calculation, accelerate convergence and obtain an acceptable computational performance. All these steps give way to an elementary and approximate method that is described in this paper.

This paper shows a method to characterize and clustering the ECPs through the use of SFs proposed by the authors and the application of *k*-means clustering technique. The proposed methodology stands out for its simplicity and efficiency.

These SFs allow the reduction of sensitivity to outliers when they are grouped. Additionally, based on unsupervised learning and in an optimize cost function the number of clusters is found, which allows obtaining patterns to understand, summarize, and categorize the electricity consumption of consumers.

This work is segmented into four sections. In section II the applied methodology is presented, in section III obtained results are analyzed. Finally, in section IV, conclusions are presented.

2 Methodology

2.1 Information gathering

The methodology requires electricity consumption data in 15-minute intervals from electricity meters.

2.2 Data characterization

2.2.1 ECPs recognition

Table 1. Time ranges.

Time Period	Description	Number of measurements taken by the electricity meter every 15 minutes
07h00 to 18h00	Ecuadorian electric commercialization time ranges for commercial and industrial consumers [11].	44
18h00 to 22h00		16
22h00 to 07h00		36
06h00 to 12h00	Time ranges are determined by the residential, commercial and industrial load curves in Ecuador.	24
12h00 to 15h00		12
15h00 to 18h00		12

Each ECP is represented by the use of *N* characteristics according to the measure data and the desired detail of the load pattern representation. In this case, the number of characteristics to represent each daily ECP is 96, since each data corresponds to the measurements average taken during every 15 minutes of an entire day.

It is necessary to define time ranges and weekdays according to the electric tariff and working patterns as a previous step to classify the ECPs [10]. Table 1 shows the time ranges and the number of measurements in each time period for commercial and industrial consumers according to Ecuadorian regulations.

2.2.2 Shape factors definition for ECPs

The SFs are normalized active power values; these are obtained in different time periods depending of time ranges previously defined in Table 1. Each factor value depends on the behavior of each consumer [12].

To obtain the SFs of ECPs, the maximum, minimum and average power values are used ($P^{MAX,\Delta_t}, P^{MIN,\Delta_t}, P^{AV,\Delta_t}$) for each defined time range [13]. A total of 23 SFs are proposed, which characterize each ECP. Table 2 summarizes the SFs proposed by the authors. The SFs are considered to determine the characteristics of ECPs in each defined time range and according to the weekday it is analyzed.

Table 2. Shape factors.

SFs	Equation	Time Period
Load Factor (*f₁*)	$f_1 = \frac{P_{average}}{P_{maximum}}$	All day
Non-uniformity coefficient (*f₂*)	$f_2 = \frac{P_{minimum}}{P_{average}}$	All day
Average Impact	$f_x = \frac{P_{average_{time\ range}}}{P_{average}}$	For every time range in Table 1
Maximum Impact	$f_y = \frac{P_{maximum_{time\ range}}}{P_{average}}$	
Minimum Impact	$f_z = \frac{P_{minimum_{time\ range}}}{P_{average}}$	
All day's maximum impact (*f₂₁*)	$f_{21} = P_{maximum}$	All day
All day's average impact (*f₂₂*)	$f_{22} = P_{average}$	
All day's minimum impact (*f₂₃*)	$f_{23} = P_{minimum}$	

Table 3. SFs' association.

Associated Factors	Dimensions	Time Period
f1-f2	2D	All day
f3-f4-f5	3D	06h00 to 12h00
f6-f7-f8	3D	12h00 to 15h00
f9-f10-f11	3D	15h00 to 18h00
f12-f13-f14	3D	07h00 to 18h00
f15-f16-17	3D	18h00 to 22h00
f18-f19-f20	3D	22h00 to 07h00
f21-f22-f23	3D	All day

There is one limitation when the direct form characteristics method is applied; this limitation takes place when visualizing the relationship between SFs in only one graphic. Therefore, there is a need to associate them in a way that they can be clearly observed as they are related (Table 3).

2.3 Clustering

The *k*-means clustering technique is a partitioning procedure based on a number of iterations [14]. The SFs are clustering in *k* mutual exclusive clusters and they return the cluster index to which each observation has been assigned. This method is applied on real observations and it creates a unique level of clusters. The *k*-means clustering technique is used since it represents a robust technique when large amounts of data are analyzed [15].

Each observation has in-space location, this allows the method clustering them in a way that, inside of each cluster the observations are as close as possible to each other and as far as possible from the other observations of the other clusters [16].

The *k*-means clustering technique calculates the centroids of each cluster, initially in an arbitrary way from the data in analysis [17], [18]. These centroids are used to classify such observations into clusters according to a selected metric. The clusters' centers are recalculated until the minimum cost function is found from the applied metric inside the method. The k-means technique uses the Euclidean distance (Ed) as the metric for data clustering [16]. The Euclidean distance between the SFs that are in the n-dimensional space is calculated. In a general form, the Euclidean distance between the SFs of $ECP_1(f_1, f_2, \dots, f_{23})$ and $ECP_2(f_1, f_2, \dots, f_{23})$, is define as:

$$E_d(\mathrm{ECP}_1, \mathrm{ECP}_2) = \sqrt{\sum_{i=1}^{23}[f_{i(ECP_1)} - f_{i(ECP_2)}]^2} \qquad (1)$$

To advance with the clustering process, it is necessary to find the maximum number of clusters for analyzed data. Then an adequate number of clusters is defined based on the use of silhouettes coefficients (SCs) of each ECP [19].

2.3.1 Maximum number of clusters

To determine the maximum number of clusters in which the SFs can be segmented, the methodology uses the Sturges' rule [20]:

$$k = 1 + \log_2(N) \qquad (2)$$

where:

- ***N*** is the number of samples (number of daily ECPs).
- ***k*** is the maximum number of clusters in which SFs can be segmented.

1) Output arguments

When the *k*-means clustering technique is applied the output arguments obtained are [12]:

- Indexes of clusters for each ECP.
- Location of the centroids according to the number of clusters specified.

2) Silhouette coefficients

The SCs are used to know the optimal number of clusters for the data. Then the SC is calculated for each ECP based on SFs. It is a measure that combines the agglomeration and separation on how close an ECP is to other ECPs in the same cluster. The SC for each ECP can vary from -1 to +1. Based on this criterion, an ECP is correctly grouped when the SC is close to 1 [14].

If majority of ECPs have a high silhouette value, the clustering is correct. The SC of each ECP calculation is based on the equation (3) [21], [22]:

$$S_i = \frac{bi - ai}{max(ai, bi)} \qquad (3)$$

where:

- $\boldsymbol{a_i}$ represents the average distance value determined from the *i*-sample to all data in the same cluster.
- $\boldsymbol{b_i}$ represents the minimum value of the average distances determined from the *i*-sample to all data in other clusters.
- $\boldsymbol{S_i}$ represents the SC for the *i*-sample.

From the equation (3), the average value of SCs is obtained according to the number of clusters with which the segmentation procedure was performed. The highest average value indicates the appropriate number of clusters to be applied [23] as shown in Figure 1.

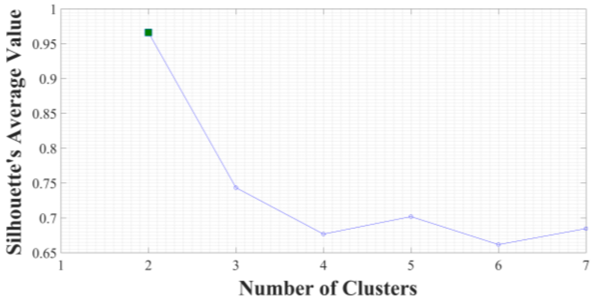

Fig. 1. Average silhouette values according to the clusters' number.

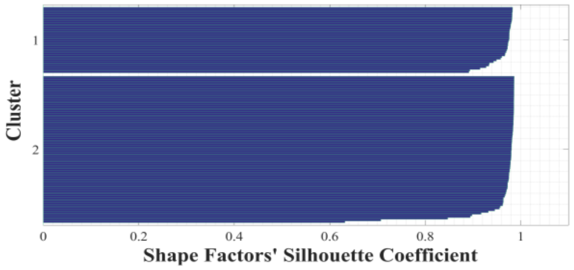

Fig. 2. SCs for each observation with 2 clusters.

3 Results analysis

The results analysis is made based on the measurements obtained from the main connection of the *Universidad Politécnica Salesiana (UPS), Sede Cuenca, Ecuador.*

3.1 Characterization of ECPs

In Figure 3, Mondays ECPs are show; while Figures 4 and 5 show the association of the SFs characterize every single ECPs in 2D and 3D correspondingly, as show in Table 3, for the reason that the 23 SFs cannot be displayed in a single figure.

In this way, Figure 4 and Figure 5 show the association of SFs in two elementary clusters. This is corroborated by determining the highest average value of the SCs.

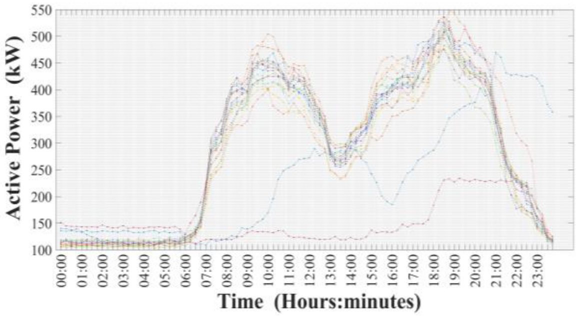

Fig. 3. Electric load profiles (Mondays).

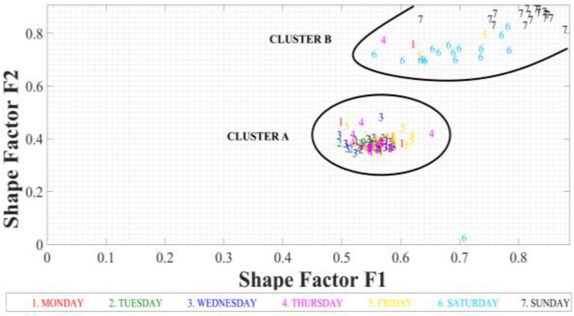

Fig. 4. SFs association (F1-F2).

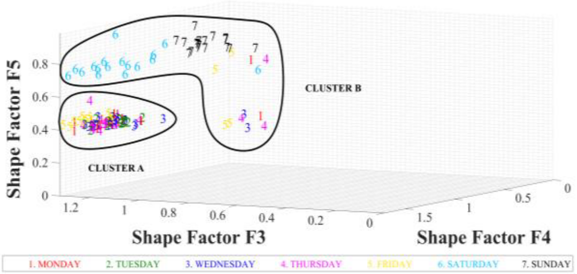

Fig. 5. SFs association (F3-F4-F5).

3.2 The k-means clustering technique application

The *k*-means application is carried out with 126 ECPs. According to the available data and the Sturges Rule, it is determined that the maximum number of possible clusters is 7.

As follows, the average values of the SCs for each possible cluster are obtained. This allows the determination of the appropriate number of clusters for the analyzed ECPs. Table 4 shows the average values of the SCs according to the number of clusters. The highest average value of the SCs shows the most appropriate number of clusters. Therefore, the 126 ECPs analyzed are segmented into 2 clusters.

Table 4. Average values of the SCs according to the number of clusters.

Clusters	Average value of the SCs
1	-
2	**0.9663**
3	0.7435
4	0.6768
5	0.6789
6	0.6618
7	0.6847

In this case, cluster A contains 87 profiles, while cluster B contains 39 profiles. These profiles represent workdays and nonworking days correspondingly.

3.3 ECPs classification

In order to carry out a correct ECPs classification, outliers are identified inside each cluster [24].

After applying unsupervised support vector machines and obtaining the best separation hyperplane between all the data inside each cluster [18, p. 53], 11 atypical profiles are detected in cluster A (Figure 6) and 10 in cluster B (Figure 7).

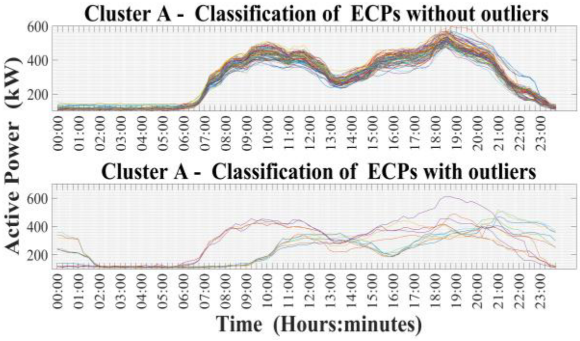

Fig. 6. Cluster A's ECPs, 76 typical profiles (above), 11 atypical profiles (below)

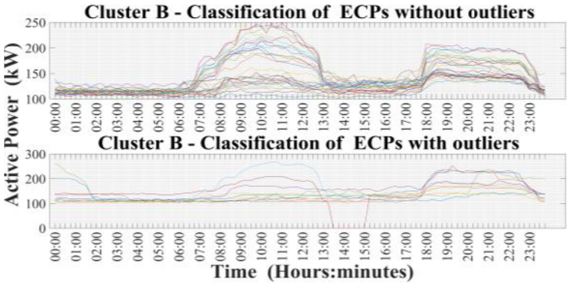

Fig. 7. Cluster B's ECPs, 29 typical profiles (above), 10 atypical profiles (below).

3.4 Obtaining patterns

The set of typical ECPs define the electrical consumption pattern. Figure 8 shows the pattern of workdays (cluster A), while Figure 9 shows the pattern of nonworking days (cluster B).

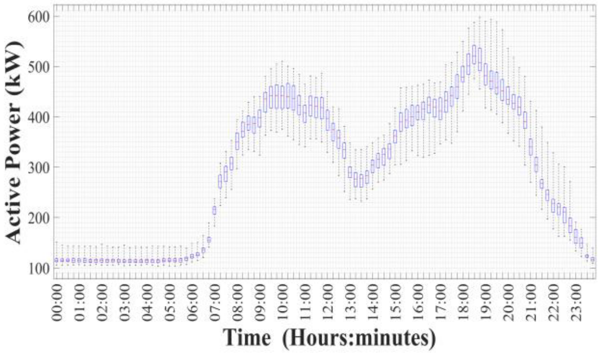

Fig. 8. Cluster A ECPs' pattern box-plot.

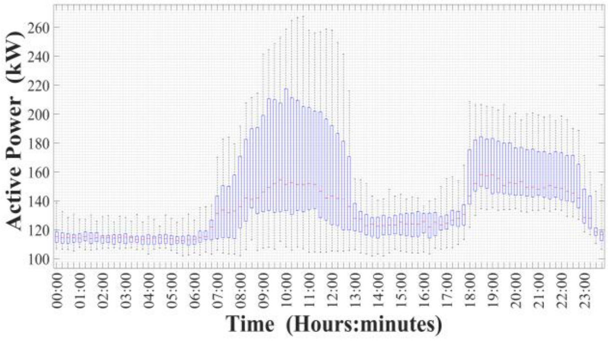

Fig. 9. Cluster B ECPs' pattern box-plot.

The electric consumption characterization allows the ECPs recognition at a measurement point. From each ECP, their respective SFs are defined, which are the basis for the clustering process. Next, atypical profiles are identified and separated. The typical profiles of each cluster represent the usual demand of the installation, that is, a consumption pattern. Consumption patterns are useful in several areas, such as forecasting and predicting demand [25], energy management, efficiency policies implementation and energy intelligence, improving the supply of tariffs, etc.

4 Conclusions

In this paper, a characterization method was applied for clustering of ECPs at UPS Cuenca, Ecuador by using SFs and the application of the *k*-means technique.

The SFs definition allows the characterization of the ECPs. This requires criteria and intuition. The *k*-means technique allows clustering the SFs in such way that the ECPs are clustered according to their similarity.

The appropriate number of clusters in which the ECPs are segmented is defined through the SCs criterion, which allows to simplify the process and improve the computational performance.

The anomalies detection process allows to obtain SCs values closer to 1 in each cluster, which makes it possible to define clearly the electric consumption characteristics. This way the understanding of the structure that they possess improves considerably.

Traditionally, most electric companies used to classify their consumers using minimum electrical parameters and some commercial regulations. This way, current electricity markets need to classify their consumers by indicators capable of characterizing their true electrical behavior.

The method allows obtaining the average daily energy consumption pattern of each cluster. This pattern is obtained from the typical profiles of each cluster and allows the determination of the kind of consumer to which it belongs; as well as defining tariffs that are clear, transparent and easy to understand, but flexible enough to follow the variations in the consumer's load pattern induced by specific rates.

If the consumption patterns are known by the consumer, they could be interested in managing the energy they use, by improving their energy performance, achieving economic benefits and decreasing their environmental footprint.

One of the limitations of the k-means method is the sensitivity to outliers.

In addition, a problem presented by the method applied in this paper is summarized in the precision with which the data can be clustered. The method works efficiently with large databases but when there are abrupt changes in the ECPs, the method can become imprecise, since SFs will be affected.

References

1. G. Chicco, R. Napoli, P. Postolache, M. Scutariu, and C. Toader, "Electric energy customer characterisation for developing dedicated market strategies," *2001 IEEE Porto Power Tech Proc.*, vol. 1, pp. 371–377, 2001.
2. G. Chicco, "Customer Behaviour and Data Analytics," *2016 Int. Conf. Expo. Electr. Power Eng.*, no. Epe, pp. 771–779, 2016.
3. J. Molina and J. García, "Técnicas de Análisis de Datos," p. 266, 2006.
4. U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From Data Mining to Knowledge Discovery in Databases," *AI Mag.*, pp. 37–54, 1996.
5. G. Chicco, R. Napoli, P. Postolache, M. Scutariu, and C. Toader, "Customer Characterization Options for Improving the Tariff Offer," *IEEE Power Eng. Rev.*, vol. 22, no. 11, p. 60, 2002.
6. G. Chicco, "Overview and performance assessment of the clustering methods for electrical load pattern grouping," *Energy*, vol. 42, no. 1, pp. 68–80, 2012.
7. G. Chicco, R. Napoli, and F. Piglione, "A Review of Concepts and Techniques for Emergent Customer Categorisation," pp. 51–58, 2002.
8. R. O. Duda, P. E. Hart, and D. G. Stork, "Pattern Classification," *New York: John Wiley, Section*. p. 654, 2000.

9. J. Valente de Oliveira and W. Pedrycz, *Advances in Fuzzy Clustering and its Applications Advances in Fuzzy Clustering and its Applications*. 2007.
10. X. Serrano-Guerrero, G. Escrivá-Escrivá, and C. Roldán-Blay, "Statistical Methodology to Assess Changes in the Electrical Consumption Profile of Buildings," *Energy Build.*, vol. 164, pp. 99–108, 2018.
11. Consejo Nacional de Electricidad, "Estudio y gestión de la demanda eléctrica," *Plan Maest. Electrif. 2013-2022*, vol. 2, pp. 41–42, 2013.
12. G. Chicco, R. Napoli, and F. Piglione, "Application of clustering techniques to load pattern-based electricity customer classification," no. June, pp. 6–9, 2005.
13. M. Piao, J. B. Lee, H. S. Shon, E. J. Cha, K. Ah Kim, and K. H. Ryu, "Identification of temporal interval relation of frequent patterns during incremental phase," *Legacy*, pp. 497–502, 2011.
14. G. K. Gupta, *Introduction To Data Mining With Case Studies*. PHI Learning Pvt. Ltd., 2014.
15. G. Chicco, R. Napoli, and F. Piglione, "Application of Clustering Algorithms and Self Organising Maps to Classify Electricity Customers," *2003 IEEE Bol. PowerTech - Conf. Proc.*, vol. 1, pp. 373–379, 2003.
16. MathWorks, "k-Means Clustering," 2017. [Online]. Available: https://www.mathworks.com/help/stats/k-means-clustering.html.
17. J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. 2012.
18. S. M. H. Jansen, "Customer Segmentation and Customer Profiling for a Mobile Telecommunications Company Based on Usage Behavior," *Proc. - 3rd Int. Conf. Data Min. Intell. Inf. Technol. Appl. ICMIA 2011*, no. July, pp. 308–313, 2007.
19. P.-N. Tan, M. Steinbach, A. Karpatne, and V. Kumar, *Introduction to Data Mining*. 2012.
20. J. J. González, "El histograma con la TI-92: optimización de clases," *Revista de Didáctica de las Matemáticas*, vol. 61. pp. 67–72, 2005.
21. P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, no. C, pp. 53–65, 1987.
22. F. Scarlatache, G. Grigora, G. Chicco, and G. Câr, "Using k-Means Clustering Method in Determination of the Optimal Placement of Distributed Generation Sources in Electrical Distribution Systems," *IEEE Optim. Electr. Electron. Equip. (OPTIM), 2012 13th Int. Conf.*, pp. 953–958, 2012.
23. MathWorks, "silhouette," 2017. [Online]. Available: https://www.mathworks.com/help/stats/silhouette.html?s_tid=doc_ta.
24. D. M. J. Tax, "One-class classification," 2001.
25. X. Serrano-Guerrero, R. Prieto-Galarza, E. Huilcatanda, J. Cabrera-Zeas, and G. Escrivá-Escrivá, "Election of Variables and Short-term Forecasting of Electricity Demand Based on Backpropagation Artificial Neural Networks," *IEEE Int. Autumn Meet. Power, Electron. Comput.*, 2017.