Electricity clustering framework for automatic classification of customer loads

Félix Biscarri *, Iñigo Monedero, Antonio García, Juan Ignacio Guerrero, Carlos León

Escuela Politécnica Superior, Electronic Department, Virgen de Africa, 7, Sevilla, CP 41011, Spain

Keywords:
Electricity consumption
Hourly demand
Load profiling
Time-series clustering
Clustering features selection
Tree classification methods

ABSTRACT

Clustering in energy markets is a top topic with high significance on expert and intelligent systems. The main impact of is paper is the proposal of a new clustering framework for the automatic classification of electricity customers' loads. An automatic selection of the clustering classification algorithm is also highlighted. Finally, new customers can be assigned to a predefined set of clusters in the classification phase. The computation time of the proposed framework is less than that of previous classification techniques, which enables the processing of a complete electric company sample in a matter of minutes on a personal computer. The high accuracy of the predicted classification results verifies the performance of the clustering technique. This classification phase is of significant assistance in interpreting the results, and the simplicity of the clustering phase is sufficient to demonstrate the quality of the complete mining framework.

1. Introduction

New technologies derived from the paradigm of Smart Grids (Tuballa & Abundo, 2016) have increased the control and monitoring of electricity consumption by customers, distribution companies, and retailers. This new scenario has led to an exponential growth in the available information concerning the grid and consumption. Thus, these technologies have led to the emergence of new services, and the increased efficiency and reliability of electricity supplies. To facilitate interaction with other systems, these new services must be able to analyse huge amounts of information in a short time (Fang et al., 2016). To achieve this goal, analysis methods and modelling designs must be constructed using big data platforms (Diamantoulakis, Kapinas, & Karagiannidis, 2015) such as Apache Hadoop (Hafen, Gibson, van Dam, & Critchlow, 2014) or Spark (Shyam, Kumar, Poornachandran, & Soman, 2015). In the current regulation model of the electricity sector, one of the main targets is to improve the performance of distribution, thus increasing the level of knowledge about demand. The most common way to evaluate energy efficiency is to evaluate the behaviour of the customers' load curve, including possible displacements in peak hours (Ferreira, de Oliveira Fontes, Cavalcante, & Marambio, 2015). Accurate knowledge of customers' consumption patterns represents a worthwhile asset for electricity providers in competitive electricity markets. Various approaches can be used to group customers that exhibit similar electricity consumption behaviour into customer classes (Chicco et al., 2004; Xu & Wunsch, 2005). Dynamic clustering can be applied (Benítez, Quijano, Díez, & Delgado, 2014; Lee, Kim, & Kim, 2011), with the focus on large-scale customers (Tsekouras, Hatziargyriou, & Dialynas, 2007; Zhang, Zhang, Lu, Feng, & Yang, 2012). The main idea is to identify customers hourly load profiles (HLPs) (Chicco, 2012; Grigoras & Scarlatache, 2014) and develop a rule set for the automatic classification of new consumers (Halkidi & Vazirgiannis, 2008; Ramos, Duarte, Duarte, & Vale, 2015). Several customer parameters, i.e. economic size, economic activity, and energy consumption, are typically used in current models (Dzobo, Alvehag, Gaunt, & Herman, 2014). In the market scenario, electricity providers have been given new degrees of freedom in defining tariff structures and rates under regulatory-imposed revenue caps (Granell, Axon, & Wallom, 2015). This requires a suitable grouping of the electricity customers into customer classes (Figueiredo, Rodrigues, Vale, & Gouveia, 2005). Other applications of load classification including the identification and correction of erroneous data and load forecasting (le Zhou, lin Yang, & Shen, 2013). Statistical techniques such as k-means (López et al., 2011), fuzzy techniques (Azadeh, Saberi, & Seraj, 2010), and frequency-domain load pattern data (Carpaneto, Chicco, Napoli, & Scutariu, 2006) have been

* Corresponding author.
E-mail addresses: fbiscarri@us.es (F. Biscarri), imonedero@us.es (I. Monedero), agarcia@us.es (A. García), juaguealo@us.es (J.I. Guerrero), cleon@us.es (C. León).

used. Although different clustering methods are used for load classification (Rasanen, Voukantsis, Niska, Karatzas, & Kolehmainen, 2010), the key requirements are some load data measuring and collection platform for automated meter reading (AMR), computing software such as MATLAB, SPSS, or R, and high-performance computers. The present study was conducted using the R software.

2. Mining framework for load classification

Load classification includes a pre-clustering phase to distinguish between different categories of customers. This first categorization of customer loads considers economic reasons. During the pre-clustering phase, the main feature is the contracted tariff, which usually determines the expected load profile. Other fields of interest are the seasonal variation in electricity consumption and the individual consumer categories: households, agriculture, industry, private services, and public services. For example, agriculture consumption is not as systematic as for the other categories, and is heavily dependent on meteorological variables such as temperature, cloud cover, and daylight hours. Consumption on workdays and non-workdays differs between months and different categories, except in certain industries where the monthly profiles are assumed to be identical. Unfortunately, individual consumer categories are not always specified in the electric company database, which means that this useful information is not available for clustering purposes. After the pre-clustering phase, a data reduction process will be performed. The sampling rate of smart meters enables 24–96 consumption data per day, i.e. a sample every hour or every 15 min. This represents a significant computation time for millions of customers, each one described with a 96-dimensional vector per day. This huge quantity of data necessitates the use of data reduction and characterization techniques. Furthermore, significant information will be preserved during the reduction process. The algorithm could be run on a big data system, such as those based on Apache Hadoop or Spark libraries. The use of these infrastructures increases the efficiency and reliability of the algorithm in large-scale databases, which decreases the computation time. Different techniques for this purpose include principal component analysis (Chicco, Napoli, & Piglione, 2006), harmonic analysis (Carpaneto, Chicco, Napoli, & Scutariu, 2006), and the wavelet representation (Mallat, 1989). López et al. (2011) proposed the daily mean power values, calculated during time-of-use pricing (usually two daily periods, named peak and valley hours). This paper presents a new data characterization that reduces the computation time with respect to the techniques mentioned above. The pre-clustering phase reduces the information on each customer to a vector composed of a few features. This vector is used as the input to the clustering process. This clustering phase includes several tasks: the selection of the clustering algorithm and the optimum number of clusters. Validation techniques (Halkidi, Batistakis, & Vazirgiannis, 2001) can be applied during this step to ensure the quality of the clustering results. The correctness of clustering algorithm results is verified using appropriate criteria and techniques. Since clustering algorithms define clusters that are not known a priori, irrespective of the clustering method, the final partitioning of data generally requires some kind of evaluation. Thus, the main output of the clustering phase is the classification of a sample of customers into clusters. In many cases, the experts in the application area have to integrate the clustering results with other experimental evidence and analysis in order to draw appropriate conclusions. In other words, electric company experts would validate and interpret the pragmatic usefulness of the clustering. The final step is the classification phase. Classification assigns new customers to a predefined set of categories or clusters. The clustering phase produces initial categories in which the values in a dataset are classified during the classification process. The classification phase is of great assistance in interpreting the results, as we show in the following sections. The simplicity of the clustering phase convinces an electric company expert of the quality of the complete mining framework (Fig. 1).

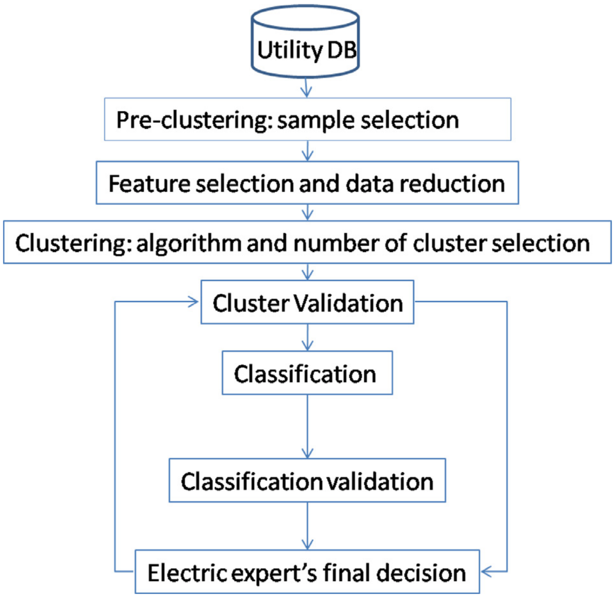
Utility DB
Pre-clustering: sample selection
Feature selection and data reduction
Clustering: algorithm and number of cluster selection
Cluster Validation
Classification
Classification validation
Electric expert's final decision

Fig. 1. Clustering and classification mining framework.

3. Pre-clustering and feature selection phases

During this phase, a sample of similar customers is selected based on their consumption and other economic criteria. The individual consumer categories are often incorrectly specified in company databases. Most customers appear as households, despite having a high contracted power. Thus, we selected a sample of customers with a certain contracted power (tariff 3.0A), a timeframe that ensures non-seasonal variations (three months), and a common climate location. These customers are located in nine adjacent villages around Seville in Andalucia, Spain. The differences between the customer environments were minimized to guarantee the homogeneity of the subsequent clustering. The sample contained a total of 218 customers. The 3.0A tariff is a time-to-use tariff for low-voltage customers (below 1 kV). There are three defined periods for pricing: peak (18–22 h), valley (0–8 h), and flat (8–18 h and 22–24 h). According to information from Spain National Commission of Energy (CNE), such customers represent approximately 2% of all electricity consumers in Spain. Feature selection and data reduction are necessary tasks. Twenty-four hourly data points per customer per day would be unmanageable in terms of computation time. Hence, researchers often use the mean daily power or the mean power during each pricing period (mean peak hours power, mean valley hours power, mean flat hours power) (López et al., 2011). Additionally, a number of studies distinguish between different months and working or non-working days (Chicco, 2012). There are two practical problems in the use of these features, one regarding the use of the mean power and another related to the use of the pricing periods. With respect to calculations based on the pricing periods, customers often vary their load profiles over the same period. Company experts prefer to divide the daytime into several periods based on the true electricity use. These periods are highly dependent on the economic activity and climate location of the consumer. Sample tests have shown that the fol-

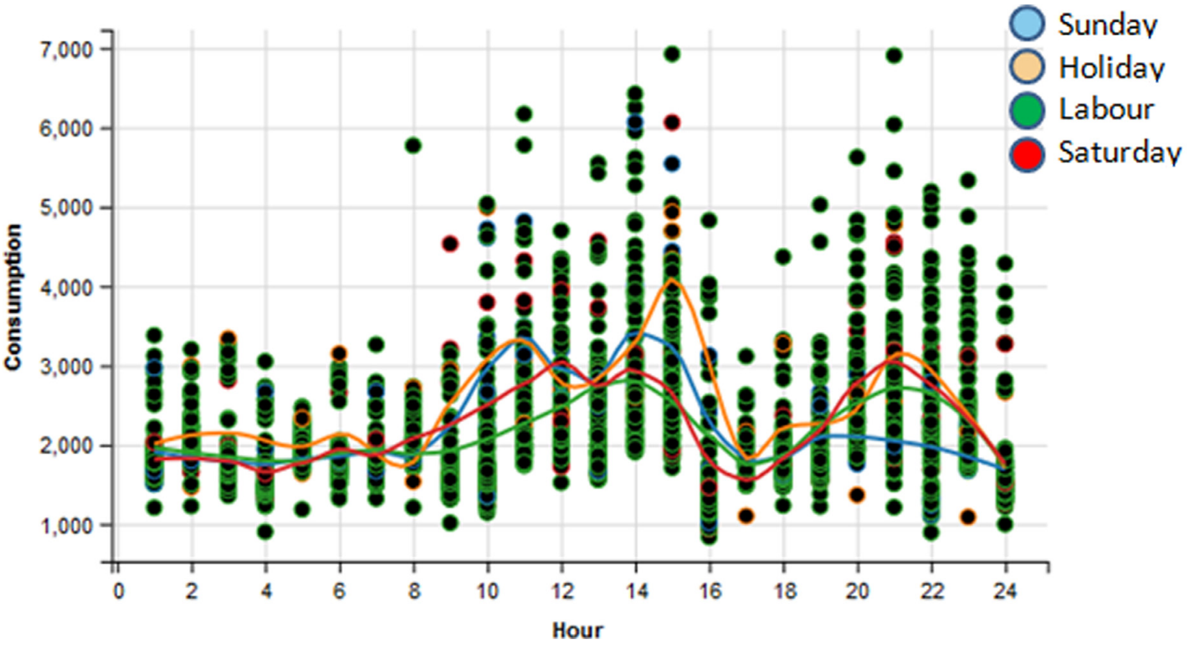

Fig. 2. LOESS curve in customer with tariff 3.0A. (For interpretation of the references to colour in the text, the reader is referred to the web version of this article.)

lowings three time periods are appropriate: 8–15 h; 16–21 h; and 22–7 h. This division considers the opening and closing times of commercial customers and the typical work day of households. Observations of a complete set of customer load curves support this decision.

With respect to the calculation of the consumption mean over several time periods, mean values mask the evolution of hourly consumption. Company experts prefer others features that allow the evolution of the load curve to be determined. We have used the number of hours of high/low consumption in the three time periods described below, as well as the count of the number of customer consumption peaks and valleys during the same time periods. Company experts can understand such features, which are also useful for clustering purposes. This set of features is fully explained in Appendix A. To distinguish peaks and valleys, a LOESS (LOcal regrESSion smoothing) curve was used. LOESS curves combine multiple regression models in a k-nearest-neighbour-based meta-model. A smoothing parameter (α) controls the flexibility of the LOESS regression function. Large values of α produce smoother functions in response to fluctuations in the data. Smaller values of α cause the regression function to follow the data more closely. However, too small a value of α is not desirable, because the regression function will eventually start to capture the random errors in the data. Useful values of the smoothing parameter typically lie in the range 0.25–0.5 for most LOESS applications. In the present study, a value of $\alpha = 0.2$ has been used. The features explained in Appendix A use these curves to identify consumption peaks and valleys and for other calculations.

Fig. 2 shows the LOESS curve for a customer on working and non-working days. Company experts could observe the consumption evolution and determine that, on working days (green curve), the customer has low consumption from 0–8 h, a maximum consumption at approximately 14 h, a minimum at 17 h, and a new peak at 21 h. Sundays (orange curve) exhibit low consumption in the afternoons.

The features for this customer are NMP = 2; NAP = 1; NNP = 0; NMiV = 1; NHcNh = 0; NLcHh = 6; NHcMh = 5; NLcMh = 0; and NLcAh = 0.

4. The clustering algorithm and the optimum number of clusters

There are a wide variety of clustering algorithms available in data mining software (Brock, Pihur, Datta, & Datta, 2008), such as connectivity-based clustering (hierarchical clustering), centroid-based clustering, distribution-based clustering, and density-based clustering. The proposed framework tests a complete set of clustering algorithms with different numbers of clusters and different validation measures, and selects the most suitable solution. A brief description of each clustering method is given below.

Hierarchical clustering is an agglomerative and hierarchical clustering algorithm that yields a dendrogram that can be cut at a chosen height to produce the desired number of clusters. Each observation is initially placed in its own cluster, and the clusters are successively joined together in order of their closeness, as determined by a dissimilarity matrix.

K-means is an iterative method that minimizes the within-class sum of squares for a given number of clusters. Clustering is strongly dependent of the initial guess for the cluster centres. Each observation is placed in the cluster to which it is closest, and the centres are updated until they remain stationary.

Diana is a hierarchical algorithm that starts with all observations in a single cluster and successively divides the observations until there are a series of clusters containing only a single observation. This algorithm uses a hierarchical divisive clustering approach.

PAM is similar to k-means, but admits the use of other dissimilarity metrics besides the Euclidean distance.

Clara is a sampling-based algorithm that implements PAM on a number of sub-datasets, which is useful when the number of observations is relatively large.

Fanny is a fuzzy-clustering algorithm in which each observation can have a partial membership in each cluster. Each observation is assigned to the cluster for which it has the highest membership value.

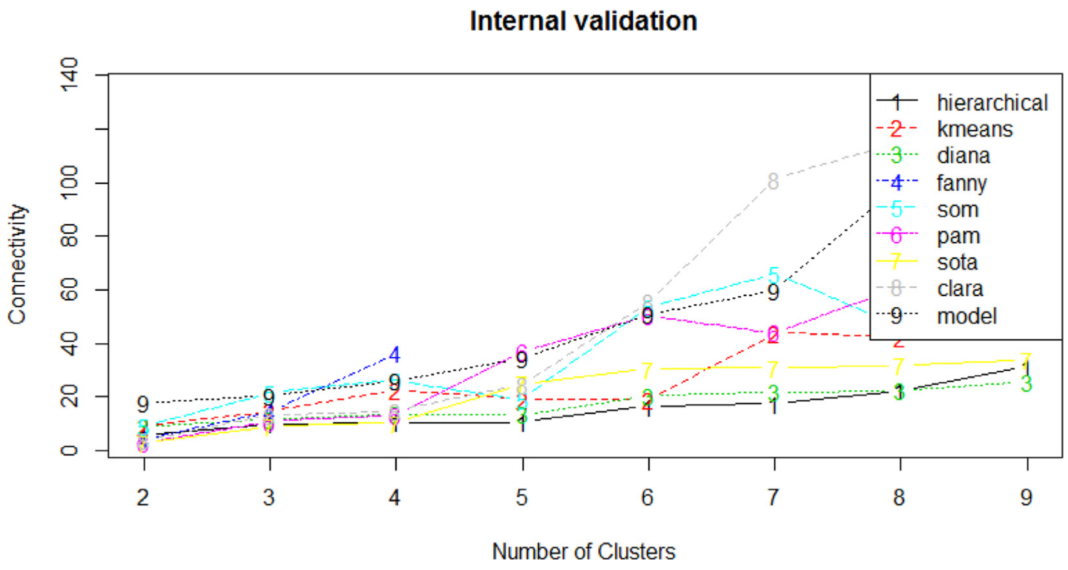

Fig. 3. Internal validation: connectivity.

Table 1
Optimal scores in validation tests.

	Score	Method	Clusters
Connectivity	2.8611	PAM	2
Dunn	0.3121	Diana	9
Silhouette	0.6217	SOM	3

SOM is a neural network unsupervised learning technique that allows maps and visualizes high-dimensional data in two dimensions.

Model-based clustering is a distribution-based technique in which a mixture of Gaussian distributions is fitted to the data. Each mixture is a cluster, and the group membership is estimated using a maximum likelihood algorithm.

SOTA is an unsupervised algorithm with a divisive tree structure.

The validation measures take only the dataset and the clustering partition as input, and use intrinsic information within the data to assess the quality of the clustering. We select measures that reflect the compactness, connectedness, and separation of the cluster partitions (internal validation indices). Connectedness is a local concept of clustering based on the idea that neighbouring data items should share the same cluster, and is here measured by the connectivity index (Handl, Knowles, & Kell, 2005). Compactness assesses the cluster homogeneity, usually by looking at the intra-cluster variance, whereas separation quantifies the degree of separation between clusters (usually by measuring the distance between cluster centroids). Because compactness and separation demonstrate opposing trends (compactness increases with the number of clusters but separation decreases), popular methods combine the two measures into a single score. The Dunn index (Dunn, 1974) and silhouette width (Rousseeuw, 1987) are examples of nonlinear combinations of the compactness and separation. Several tests were completed using the R software. The clustering algorithms and validation measures (with 2–9 clusters) generated the optimal scores presented in Table 1. Three different solutions were obtained using the three validation indexes. This is a typical result: different methods produce different numbers of clusters selected by different validation indexes.

These results should be examined carefully. With respect to the connectivity index (where smaller numbers are better), Fig. 3 shows the complete test results. Low cluster numbers (2–5 clusters) give a positive outcome. Diana and the hierarchical algorithms performed well with 2–9 clusters, whereas the PAM algorithm attained a local minimum with two clusters.

Higher values of the Dunn index are better. This index indicates good performance with four clusters, particularly with the hierarchical and Diana algorithms (Fig. 4). K-means with four clusters also performed well. The Diana algorithm gives the best selection with seven clusters.

Similarly, higher values of the silhouette index are better. This index suggests that most algorithms achieved optimal performance with three clusters, although SOM and model-based clustering did not perform well in this test (Fig. 5).

In summary, the three validity indexes indicate that the hierarchical algorithm achieves good performance with a low number of clusters. Company experts consider that this data sample can be correctly represented by a low number of clusters. Based on the expertsreports, the nine clusters that the Diana algorithm selected for the Dunn index were excessive for this selection.

5. Cluster validation by company experts and classification

Hierarchical clustering consistently performs well for most of the validation measures. Here, we extract the results from hierarchical clustering, plot the resulting dendrogram, and view the observations that are grouped together at various levels of the topology. As suggested by the company experts, two clusters is not a desirable result, as only night and day will be differentiated. Thus, three clusters were formed. The dendrogram is plotted in Fig. 6, and the mean number of features per cluster is listed in Table 2. Further inspection of the results was conducted by a subject matter expert.

Finally, the plots of the mean LOESS curves per group allow us to understand the cluster significance. Experimental outcomes shown insightful implications of customers' behaviour. Thus, the first cluster (Cluster 1, 28 customers) corresponds to night-time consumption customer. Therefore may involve that most of this energy consumption is used for lighting, and as such there is no essential difference between working and non-working days, as shown in Fig. 7.

The second and third clusters (134 customers and 56 customers, respectively) correspond to daytime consumption groups. They differ in the number of morning consumption peaks and the number

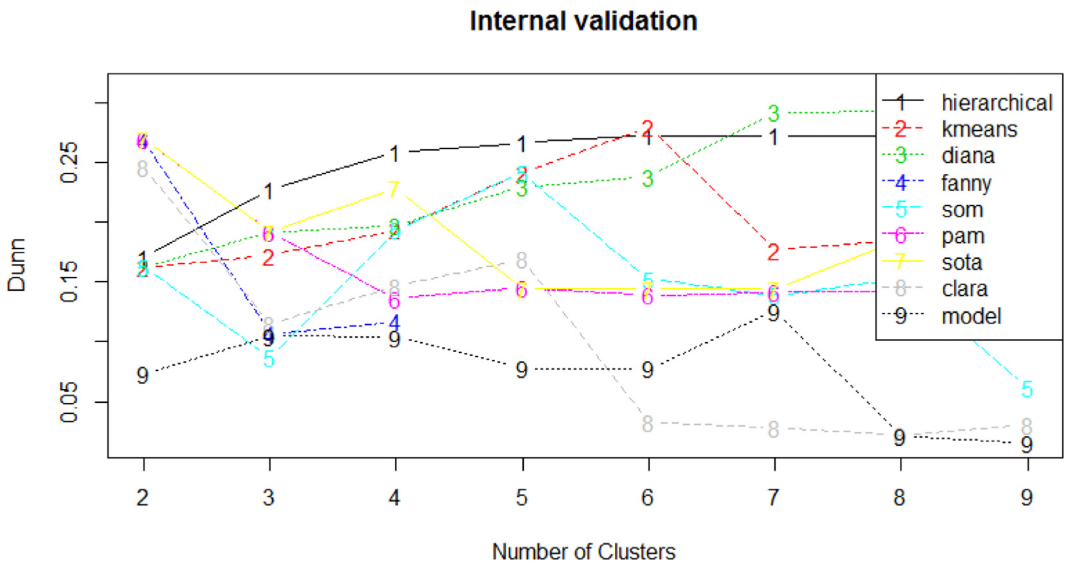

Fig. 4. Internal validation: Dunn index.

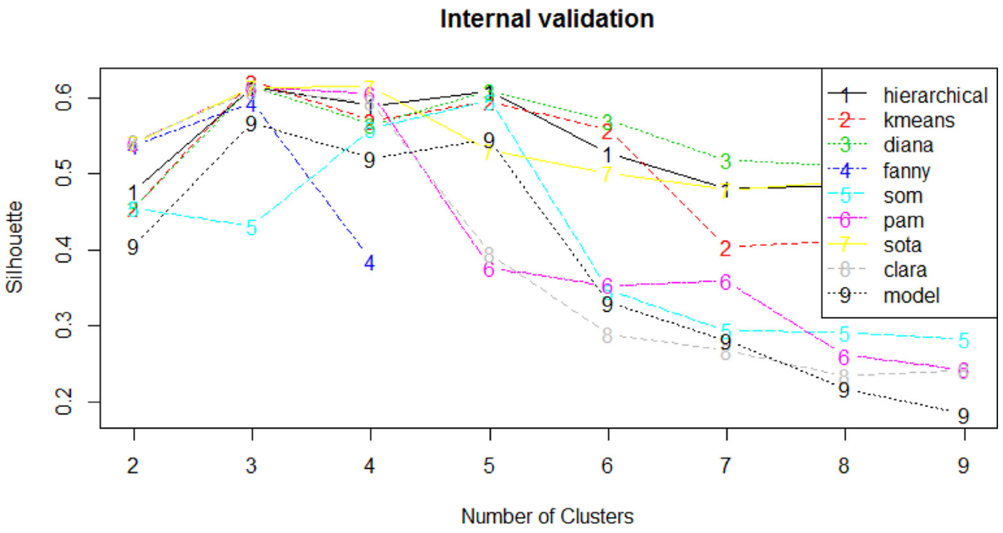

Fig. 5. Internal validation: Silouette index.

Table 2
Mean features per cluster & group, 3 clusters.

Group	NMoP	NNP	NAP	NMiV	NNH	NLcNh	NLcMh	NLcAh
1	0.79	1.65	0.59	0.69	7.59	0.14	2.65	0.17
2	1.59	1.16	0.92	0.03	0.75	0.16	0.07	0.02
3	1.70	0.12	0.96	0.48	0.07	7.32	0.28	0.86

of hours with high consumption. Cluster 3 exhibits low consumption on Sundays, Saturday evenings, and holidays (Figs. 8 and 9). Thus, company expert suppose that cluster 2 corresponds to a domestic customers, and cluster 3 involves a commercial consumption.

For the purpose of classification, classification and regression trees (CARTs) can be used as an alternative to logistic regression. Recursive partitioning, which builds the model in a forward stepwise search, is a popular CART method that can be used to compute the probability of being in any hierarchical group. Although this approach is known to be an efficient heuristic, the results of recursive tree methods are only locally optimal, as splits are chosen to maximize homogeneity at the next step only. An alternative means of searching the parameter space of the trees is to use global optimization methods such as evolutionary algorithms. Such approaches can be used to reduce the a priori bias. In the proposed framework, globally optimal CARTs are implemented, and the results are shown in Fig. 10.

Nodes 5 & 6 include 28 customers, all of whom belong to cluster 1. The associated rule is: $NLcNH < 5\,\&\,(NLcMH \geq 5\ or\ NLcMH < 5\,\&\,NcNH \geq 7)$. Node 4 includes 134 customers, all of whom belong to cluster 2. The associated rule is: $NLcNH < 5\,\&\,NLcMH < 5\,\&\,NcNH < 7$. Node 7 include 56 customers, all of whom belong to cluster 3. The associated rule is: $NLcNH \geq 5$. As shown, the classification method employed here correctly classifies the customers and clusters.

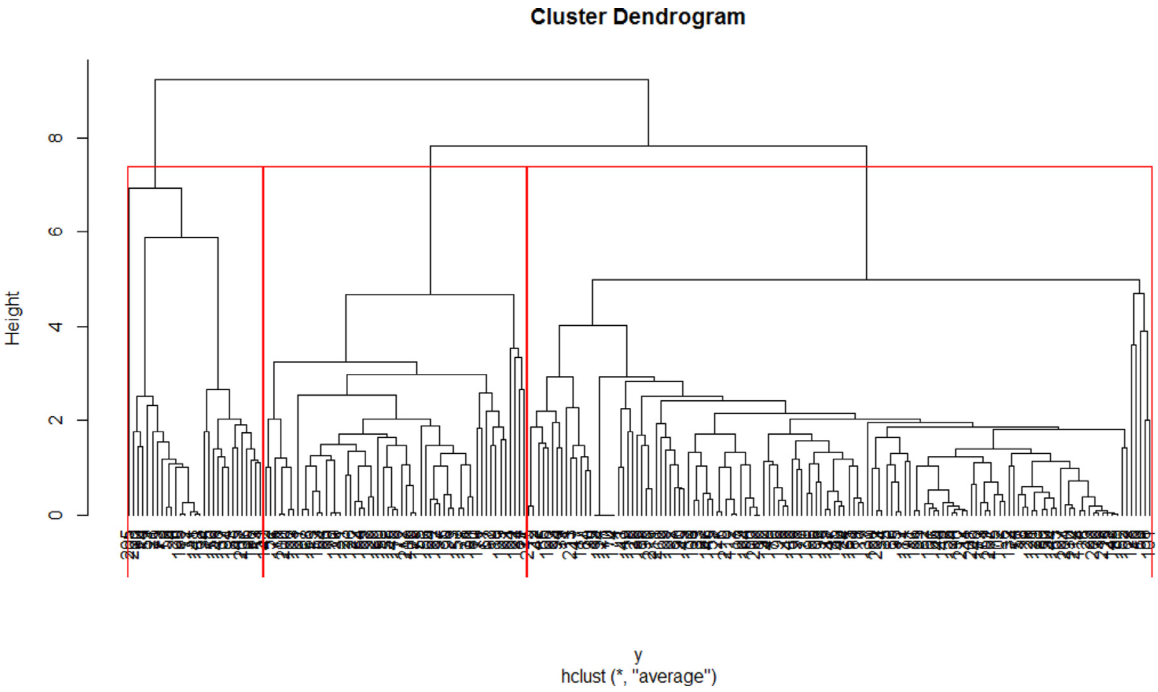

Fig. 6. Cluster dendogram, for three clusters.

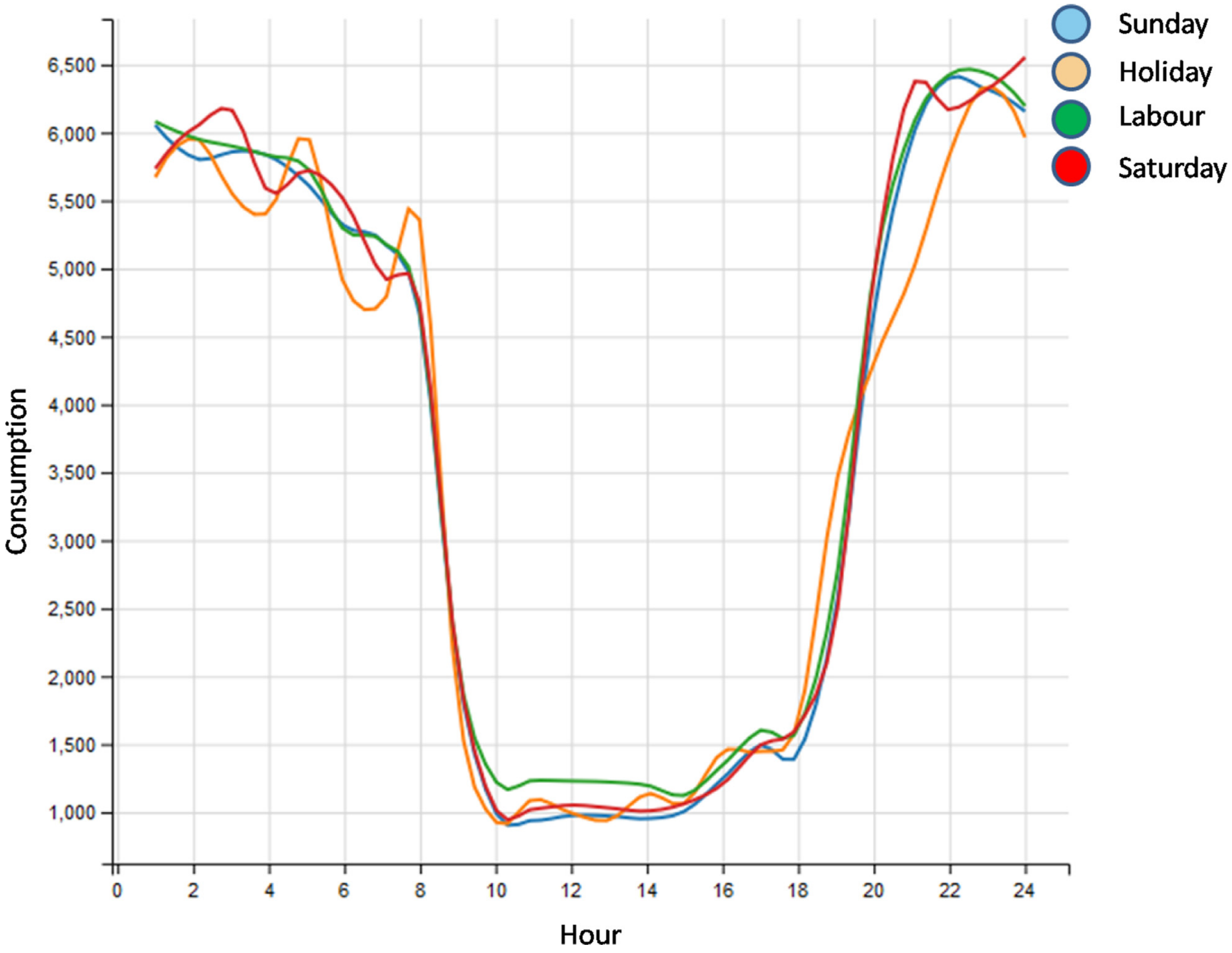

Fig. 7. Cluster 1. Mean LOESS curve, working and no-working days.

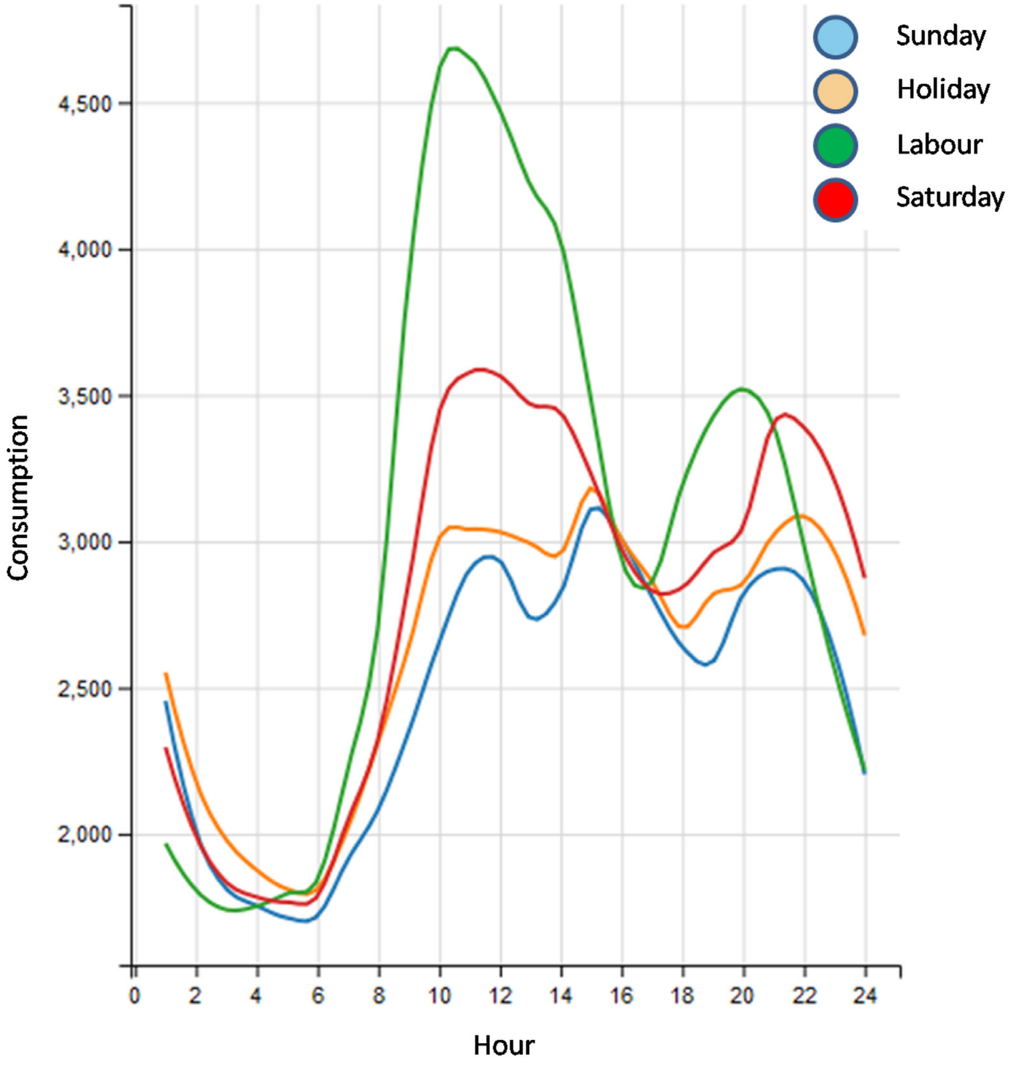

Fig. 8. Cluster 2. Mean LOESS curve, working and no-working days.

6. Conclusion

Clustering is an unsupervised machine learning technique widely used to:

- Understanding market strategies. E.g., Lorentz, Hilmola, Malmsten, and Srai (2016) uses clustering to detect small and medium size enterprises? (SME) manufacturing strategy.
- Improving the efficiency of a service. E.g., clustering techniques were utilized for improving the efficiency of a hospital's service and of the maintenance tasks in a clinical engineering department (Cruz, 2013).
- Discovering association in different markets. E.g., in Dias and Ramos (2014), the energy time series (crude oil, natural gas and electricity prices) are clustered into homogeneous groups based on their state dynamic.
- Customer grouping. In Ho, Ip, Lee, and Mou (2012), a robust genetic algorithm (GA) based k-means clustering algorithm is proposed in attempt to classify existing customers of the enterprise into groups with consideration of relevant attributes for the sake of obtaining desirable grouping results in an efficient manner.

Differentiate exiting customers with common features into smaller groups can serve as a piece of useful reference for decision-making. The main objective of this paper was to describe a complete framework for the automatic classification of electricity customers loads. The clustering of electricity load curves is a topic of significant interest, but several steps described in this paper have not been adequately covered in the literature. Authors usually select or propose a specific clustering classification algorithm, and compare their results with other solutions (Tsekouras, Hatziargyriou, & Dialynas, 2007; Zhang, Zhang, Lu, Feng, & Yang, 2012). This is probably the main strength of this paper: the algorithm selection phase that does not assume any a priori solution. The reason is that the particular solution will depend on the selected sample, because different customer categories have very different consumption patterns, and no simple algorithm can identify the optimum for all mixtures of customers.

The use of a complete set of validation indexes that reflect the connectedness (connectivity index), compactness, and separation (Dunn and silhouette indexes) of the clustering partition is another innovative step in the proposed framework. The interpretation of partial results would be similar to that given by a company expert. In this sense, local optima of the indexes would be avoided. The evolution of all indexes should be carefully studied and the expertise of electricity companies considered.

This research method is actually in the test phase. Real data from a medium-size electricity company located in the south of

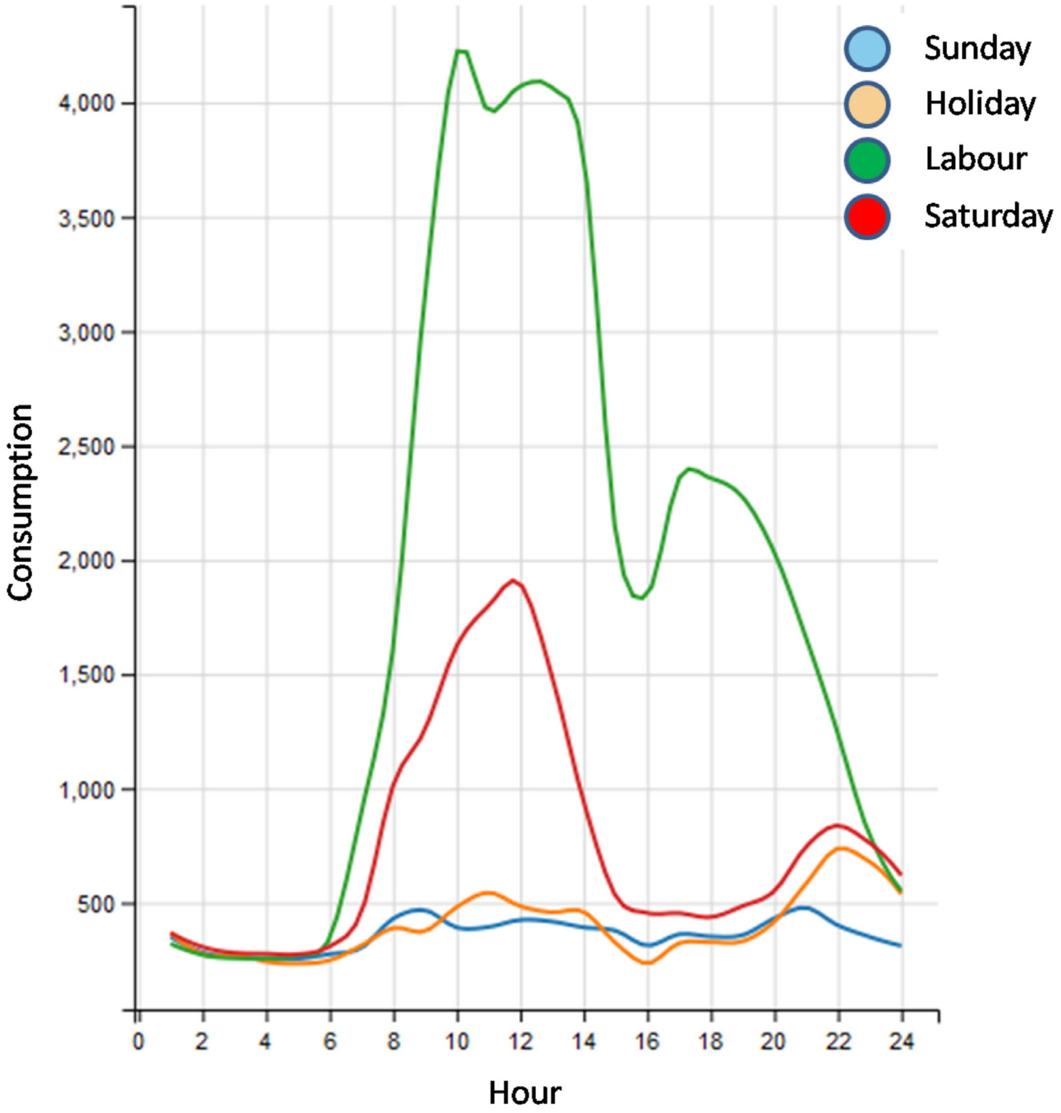

Fig. 9. Cluster 3. Mean LOESS curve, working and no-working days.

Spain have been used. Results are promising, but more practical test are needed.

The proposed framework also includes a classification phase that allows new customers to be assigned to a predefined cluster. This phase is useful for interpreting the results and convincing company experts of the quality and pragmatism of the mining framework. The high accuracy of our results supports the proposed framework. The reduction of any a priori bias can be implemented using an evolutionary algorithm that provides a global optimization method.

Future endeavours of this work are directed toward the inclusion of this work as a part of a data analysis framework for the telemetry and management of electricity distribution companies. In future research, this clustering framework will be applied to generate the input to a non-technical losses (NTLs) model in power systems analysis. NTLs are caused by actions external to the power system, such as electricity theft, non-payment by customers, or errors in accounting. NTL detection requires supervised data mining and learning, and the clustering results described in this paper will provide new and interesting information.

Future endeavours of this work are directed toward the inclusion of this work as a part of a data analysis framework for the telemetry and management of electricity distribution companies. In future research, this clustering framework will be applied:

- To generate the input to a non-technical losses (NTLs) model in power systems analysis. NTLs are caused by actions external to the power system, such as electricity theft, non-payment by customers, or errors in accounting. NTL detection requires supervised data mining and learning, and the clustering results described in this paper will provide new and interesting information.
- To investigate the economical aspects related to the possible tariff diversification for the various customer classes. Load profiles would be used for providing suggestions on possible market strategies seen from the point of view of the electricity utility.
- To differentiate exiting electric customers with common features into smaller groups. The acquired knowledge will be a useful reference for decision-making.
- To contribute to the power grid information. The future of power grids is expected to involve an increasing level of intelligence and integration of new information and communication technologies in every aspect of the electricity system, from

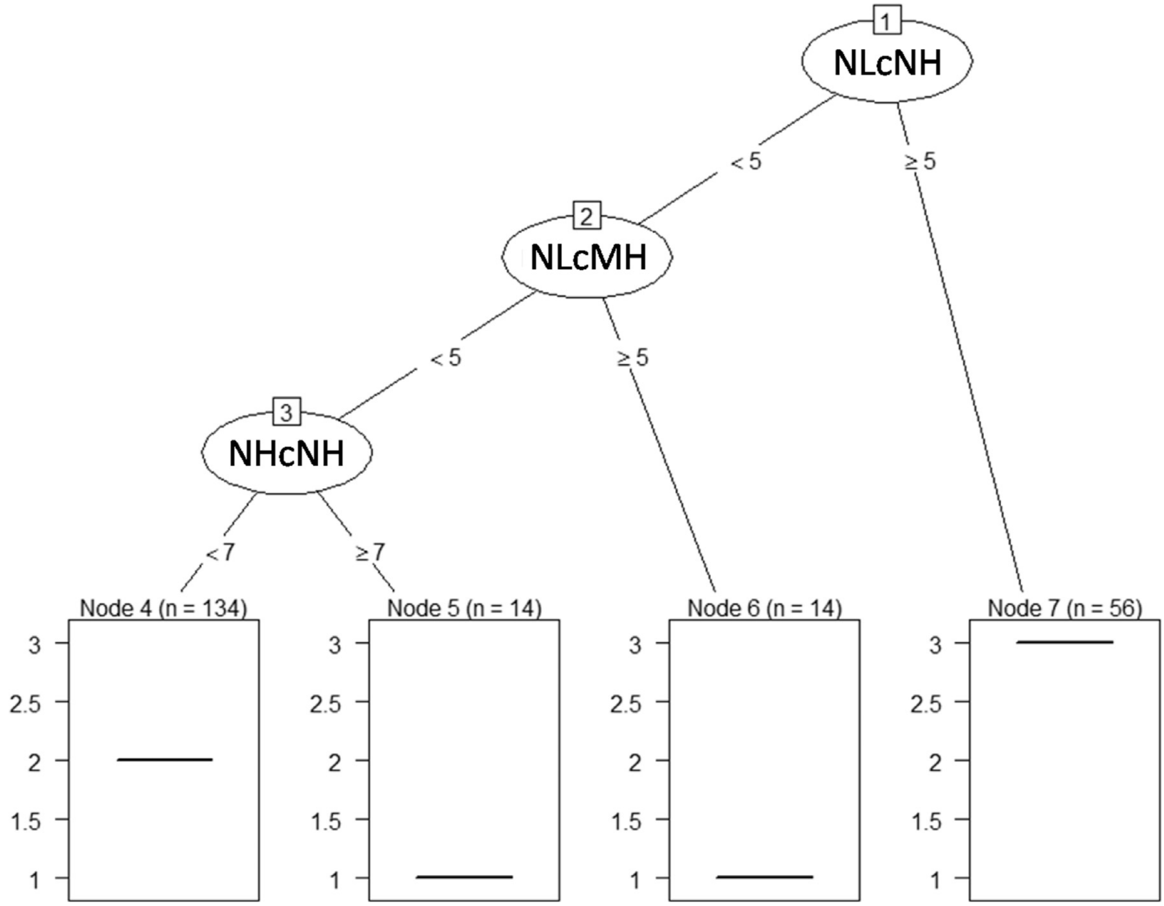

Fig. 10. Classification Tree.

demand-side devices to wide-scale distributed generation to a variety of energy markets (Coll-Mayor, Paget, & Lightner, 2007).

Acknowledgements

The work supporting this paper was funded by the projects "SIIAM: Sistema Inteligente Inalámbrico para Análisis y Monitorización de líneas de tensión subterráneas en Smart Grids" (TEC2013-40767-R, Proyects R&D+I - Spanish State Program for Research, Development and Innovation. Wireless Intelligent System for the Analysis and Monitoring of voltage underground lines in Smart Grids) and "TLG2: Plataforma de análisis de datos de Telemedida y Telegestión en Distribuidoras Eléctricas" (IDI-20150044, Data Analysis Framework for the Telemetry and Management in Electricity Distribution Companies). The authors would like to thank Isotrol Company, especially J. García Franquelo, for their contribution regarding smart meter data; Jesús Váquez and Alicia Valverde for their assistance with the smart meter dataset handling.

We sincerely thank reviewers for constructive criticisms and valuable comments. Accordingly, the revised manuscript has been improved with new information about strengths, impact, significance and novelty of the paper.

Appendix A. Features for clustering

After data normalization, in which each registered hourly data point is divided by the maximum consumption during the sample period for that customer, we obtain LOESS curves for each customer. The data features used for the clustering process are as follows:

A1. Features that count peaks or valleys throughout the specified time frame

Number of Morning Peaks (NMoP). Count the number of morning peaks on the customer's LOESS curve (from 8 to 13 h). Peaks are defined as points greater than 10% of the minimum normalized consumption registered value ($\geq$ 0.1). Number of Afternoon Peaks (NAP). Count the number of afternoon peaks on the customer's LOESS curve (from 16 to 21 h). Peaks are defined as points greater than 10% of the minimum normalized consumption registered value ($\geq$ 0.1).

Number of Night Peaks (NNP). Count the number of night peaks on the customer's LOESS curve (from 22 to 7 h). Peaks are defined as points greater than 10% of the minimum consumption registered value ($\geq$ 0.1). Number of Midday Valleys (NMiV). Count the number of midday valleys. This feature detects low consumption during midday relaxation hours (from 14 to 20 h), usually in commercial and industrial customers. Peaks are defined as points lower than 10% of the minimum consumption registered value ($\leq$ 0.1).

A2. Features that count high or low consumption hours throughout the specified time frame

Number of High consumption Night hours (NHcNh). Count the number of night hours with high consumption (from 24 to 7 h). High consumption hours are those for which the customer's consumption is greater than 60% of the maximum registered consumption ($\geq$ 0.6). Number of Low consumption Night hours (NLcNh). Count the number of night hours with low consumption (from 24 to 7 h). Low consumption hours are those for which the customer's consumption is lower than 10% of the maximum registered consumption ($\leq$ 0.1). Number of High consumption Morning hours (NHcMh). Count the number of morning hours with high consumption (from 10 to 14 h). High consumption hours are those for which the customer's consumption is greater than 60% of the maximum registered consumption ($\geq$ 0.6). Number of Low consumption Morning hours (NLcMh). Count the number of morning hours with low consumption (from 10 to 14 h). Low consumption hours are those for which the customer's consumption is lower than 10% of the maximum registered consumption ($\leq$ 0.1). Number of Low consumption Afternoon hours (NLcAh). Count the number of afternoon hours with low consumption (from 19 to 21 h). Low consumption hours are those for which the customer's consumption is lower than 10% of the maximum registered consumption ($\leq$ 0.1).

References

Azadeh, A., Saberi, M., & Seraj, O. (2010). An integrated fuzzy regression algorithm for energy consumption estimation with non-stationary data: A case study of Iran. *Energy, 35*(6), 2351–2366.

Benítez, I., Quijano, A., Díez, J.-L., & Delgado, I. (2014). Dynamic clustering segmentation applied to load profiles of energy consumption from spanish customers. *International Journal of Electrical Power & Energy Systems, 55*, 437–448.

Brock, G., Pihur, V., Datta, S., & Datta, S. (2008). clvalid: An R package for cluster validation. *Journal of Statistical Software, 25*(1), 1–22.

Carpaneto, E., Chicco, G., Napoli, R., & Scutariu, M. (2006). Electricity customer classification using frequency-domain load pattern data. *International Journal of Electrical Power & Energy Systems, 28*(1), 13–20.

Chicco, G. (2012). Overview and performance assessment of the clustering methods for electrical load pattern grouping. *Energy, 42*(1), 68–80.

Chicco, G., Napoli, R., & Piglione, F. (2006). Comparisons among clustering techniques for electricity customer classification. *IEEE Transactions on Power Systems, 21*(2), 933–940.

Chicco, G., Napoli, R., Piglione, F., Postolache, P., Scutariu, M., & Toader, C. (2004). Load pattern-based classification of electricity customers. *IEEE Transactions on Power Systems, 19*(2), 1232–1239.

Coll-Mayor, D., Paget, M., & Lightner, E. (2007). Future intelligent power grids: Analysis of the vision in the european union and the United States. *Energy Policy, 35*(4), 2453–2465.

Cruz, A. M. (2013). Evaluating record history of medical devices using association discovery and clustering techniques. *Expert Systems with Applications, 40*(13), 5292–5305.

Diamantoulakis, P. D., Kapinas, V. M., & Karagiannidis, G. K. (2015). Big data analytics for dynamic energy management in smart grids. *Big Data Research, 2*(3), 94–101.

Dias, J. G., & Ramos, S. B. (2014). Dynamic clustering of energy markets: An extended hidden markov approach. *Expert Systems with Applications, 41*(17), 7722–7729.

Dunn, J. C. (1974). Well-separated clusters and optimal fuzzy partitions. *Journal of Cybernetics, 4*(1), 95–104.

Dzobo, O., Alvehag, K., Gaunt, C., & Herman, R. (2014). Multi-dimensional customer segmentation model for power system reliability-worth analysis. *International Journal of Electrical Power & Energy Systems, 62*, 532–539.

Fang, B., Yin, X., Tan, Y., Li, C., Gao, Y., Cao, Y., & Li, J. (2016). The contributions of cloud technologies to smart grid. *Renewable and Sustainable Energy Reviews, 59*, 1326–1331.

Ferreira, A. M. S., de Oliveira Fontes, C. H., Cavalcante, C. A. M. T., & Marambio, J. E. S. (2015). Pattern recognition as a tool to support decision making in the management of the electric sector. part ii: A new method based on clustering of multivariate time series. *International Journal of Electrical Power & Energy Systems, 67*, 613–626.

Figueiredo, V., Rodrigues, F., Vale, Z., & Gouveia, J. B. (2005). An electric energy consumer characterization framework based on data mining techniques. *IEEE Transactions on Power Systems, 20*(2), 596–602.

Granell, R., Axon, C. J., & Wallom, D. C. (2015). Clustering disaggregated load profiles using a dirichlet process mixture model. *Energy Conversion and Management, 92*, 507–516.

Grigoras, G., & Scarlatache, F. (2014). Knowlegde extraction from smart meters for consumer classification. In *Electrical and power engineering (EPE), 2014 international conference and exposition* (pp. 978–982).

Hafen, R., Gibson, T., van Dam, K. K., & Critchlow, T. (2014). Chapter 1- power grid data analysis with R and hadoop. In Y. Zhao, & Y. Cen (Eds.), *Data mining applications with R* (pp. 1–34). Boston: Academic Press.

Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2001). On clustering validation techniques. *Journal od Intelligent Information Systems, 17*(2), 107–145.

Halkidi, M., & Vazirgiannis, M. (2008). A density-based cluster validity approach using multi-representatives. *Pattern Recognition Letters, 29*(6), 773–786.

Handl, J., Knowles, J., & Kell, D. B. (2005). Computational cluster validation in post-genomic data analysis. *Bioinformatics, 21*(15), 3201–3212.

Ho, G., Ip, W., Lee, C., & Mou, W. (2012). Customer grouping for better resources allocation using GA based clustering technique. *Expert Systems with Applications, 39*(2), 1979–1987.

Lee, S., Kim, G., & Kim, S. (2011). Self-adaptive and dynamic clustering for online anomaly detection. *Expert Systems with Applications, 38*(12), 14891–14898.

López, J. J., Aguado, J. A., Martín, F., noz, F. M., Rodríguez, A., & Ruiz, J. E. (2011). Hopfield?k-means clustering algorithm: A proposal for the segmentation of electricity customers. *Electric Power Systems Research, 81*(2), 716–724.

Lorentz, H., Hilmola, O.-P., Malmsten, J., & Srai, J. S. (2016). Cluster analysis application for understanding SME manufacturing strategies. *Expert Systems with Applications, 66*, 176–188.

Mallat, S. G. (1989). A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 11*(7), 674–693.

Ramos, S., Duarte, J. M., Duarte, F. J., & Vale, Z. (2015). A data-mining-based methodology to support MV electricity customers' characterization. *Energy and Buildings, 91*, 16–25.

Rasanen, T., Voukantsis, D., Niska, H., Karatzas, K., & Kolehmainen, M. (2010). Data-based method for creating electricity use load profiles using large amount of customer-specific hourly measured electricity use data. *Applied Energy, 87*(11), 3538–3545.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics, 20*, 53–65.

Shyam, Ganesh, B., Kumar, S., Poornachandran, P., & Soman (2015). Apache spark a big data analytics platform for smart grid. *Procedia Technology, 21*, 171–178.

Tsekouras, G. J., Hatziargyriou, N. D., & Dialynas, E. N. (2007). Two-stage pattern recognition of load curves for classification of electricity customers. *IEEE Transactions on Power Systems, 22*(3), 1120–1128.

Tuballa, M. L., & Abundo, M. L. (2016). A review of the development of smart grid technologies. *Renewable and Sustainable Energy Reviews, 59*, 710–725.

Xu, R., & Wunsch, D. (2005). Survey of clustering algorithms. *IEEE Transactions on Neural Networks, 16*(3), 645–678.

Zhang, T., Zhang, G., Lu, J., Feng, X., & Yang, W. (2012). A new index and classification approach for load pattern analysis of large electricity customers. *IEEE Transactions on Power Systems, 27*(1), 153–160.

le Zhou, K., lin Yang, S., & Shen, C. (2013). A review of electric load classification in smart grid environment. *Renewable and Sustainable Energy Reviews, 24*, 103–110.