Electrical Power and Energy Systems 53 (2013) 824–831

Contents lists available at SciVerse ScienceDirect

Electrical Power and Energy Systems

journal homepage: www.elsevier.com/locate/ijepes

A new method for pattern recognition in load profiles to support decision-making in the management of the electric sector

Adonias M.S. Ferreira [a,*], Carlos A.M.T. Cavalcante [a], Cristiano H.O. Fontes [a], Jorge E.S. Marambio [b,1]

[a] Program of Industrial Engineering, Polytechnic School, Federal University of Bahia, Rua Professor Aristides Novis, 2, Federação, CEP 40210-630 Salvador, BA, Brazil
[b] Norsul Engenharia LTDA, Av. Tancredo Neves 1632, S1802, Edf. Salvador Trade Center, Torre Sul, Caminho das Árvores, CEP 41820 020 Salvador, BA, Brazil

ARTICLE INFO

Article history:
Received 25 July 2012
Received in revised form 29 May 2013
Accepted 1 June 2013

Keywords:
Load profiles
Clustering
Pattern recognition
Electric sector

ABSTRACT

This work presents a method for the selection, typification and clustering of load curves (STCL) capable of recognizing consumption patterns in the electricity sector. The algorithm comprises four steps that extract essential features from the load curve of residential users with an emphasis on their seasonal and temporal profile, among others. The method was successfully implemented and tested in the context of an energy efficiency program carried out by the Energy Company of Maranhão (Brazil). This program involved the replacement of refrigerators in low-income consumers' homes in several towns located within the state of Maranhão (Brazil). The results were compared with a well known time series clustering method already established in the literature, Fuzzy CMeans (FCM). The results reveal the viability of the STCL method in recognizing patterns and in generating conclusions coherent with the reality of the electricity sector. The proposed method is also useful to support decision-making at management level.

1. Introduction

Multivariate analysis is a powerful tool for knowledge extraction especially when applying pattern recognition techniques based on data. In this context, the analysis of energy consumption data measured in homes can identify opportunities for improvement in the load factor [1] and energy efficiency of the distribution system through specific actions by the customer [2]. Methods of Data Mining (DM) that can extract useful information from data can be used to develop decision-making tools so as to improve production systems and management technology [3–6].

Some works present the use of clustering and load curve typification (pattern recognition) methods in the electric power sector. Gerbec et al. [7] performed a load curve typification using a hierarchical clustering method highlighting the advantage of this method in choosing the appropriate number of groups. The non-hierarchical method [8] emphasizes the minimization of internal variance within a cluster and also the reduction of similarity between different groups. Geminagnani et al. [9] combined the hierarchical and non-hierarchical clustering methods to improve clustering efficiency in the recognition of different consumption patterns at the same level of tension. Zalewski [10] used fuzzy logic for clustering and load curve typification. The author performed the clustering of load profiles in order to classify substations into homogeneous groups according to consumption peak. Nizar et al. [11] combined two methods, namely, Feature Selection and Knowledge Discovery in Databases (KDD) [12,13], to obtain better patterns of load demand in a distribution system. A recent study about knowledge extraction from electric power consumer data [10] presents an overall analysis and prediction of energy consumption trends (Incremental Summarization and Pattern Characterization – ISPC).

Some studies compare the performance of various methods of typification and conclude that the fuzzy C-Means (FCM) provides the best level of cohesion and discrimination of the problems associated with clustering in load curves. From this very point of view some authors have recently highlighted the FCM method in applications involving pattern recognition (typification) in load curves [14–16].

This study proposes a new method of selection, typification, and load curve clustering (STCL) based on a systematic extraction of features. This method is capable of identifying a greater diversity in demand patterns and also represents a potential tool for the improvement of the decision-making process through better classification of heterogeneous consumer profiles in the electric power sector. The case study analyzed is an energy efficiency program [17] carried out by the Electric Company of Maranhão (Brazil), that considers, among others, the analysis of the impact of replacing refrigerators in low-income consumers' homes distributed in several towns located within the state of Maranhão (Brazil). The proposed method incorporates multiple criteria in the clustering

* Corresponding author. Tel.: +55 (71) 3203 9806; fax: +55 (71) 3203 9802.
E-mail addresses: adoniasmagdiel@ufba.br (A.M.S. Ferreira), arthurtc@ufba.br (C.A.M.T. Cavalcante), cfontes@ufba.br (C.H.O. Fontes), contato@norsulengenharia.com.br (J.E.S. Marambio).
[1] Tel./fax: +55 (71) 3342 7013.

0142-0615/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.ijepes.2013.06.001

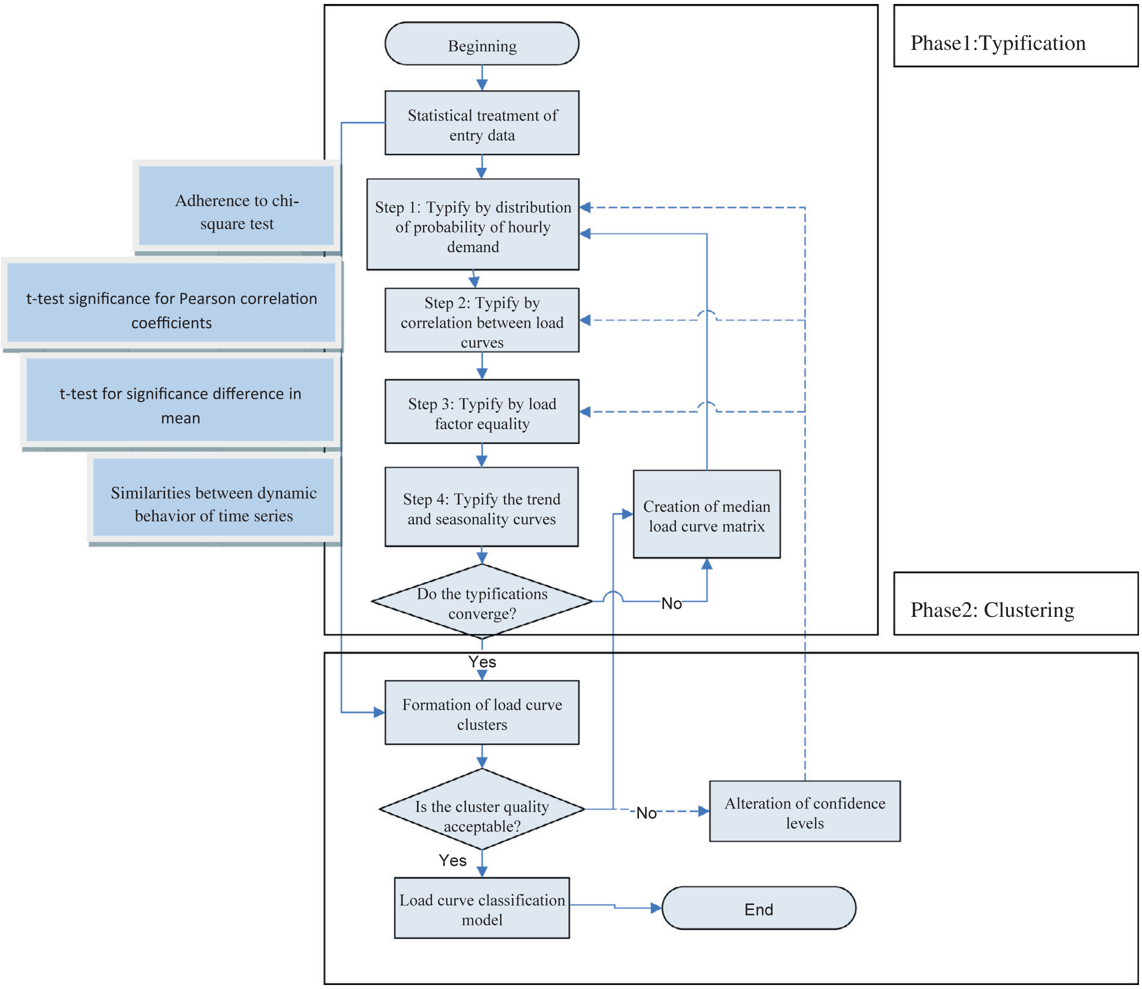
Phase1:Typification
Beginning
Statistical treatment of entry data
Adherence to chi-square test
Step 1: Typify by distribution of probability of hourly demand
t-test significance for Pearson correlation coefficients
Step 2: Typify by correlation between load curves
t-test for significance difference in mean
Step 3: Typify by load factor equality
Similarities between dynamic behavior of time series
Step 4: Typify the trend and seasonality curves
Creation of median load curve matrix
Do the typifications converge?
No
Phase2: Clustering
Yes
Formation of load curve clusters
Is the cluster quality acceptable?
No
Alteration of confidence levels
Yes
Load curve classification model
End

Fig. 1. The STCL method.

and typification of load curves unlike traditional approaches that essentially use the criterion of distance between load curves for cluster recognition. Section 2 presents the STCL method and the evaluation metrics adopted. Section 3 presents the case study and results obtained from the application of FCM and STCL methods demonstrating the ability and superiority of the latter in describing the problem.

2. The STCL method

The STCL method (Fig. 1) comprises two phases. The first carries out pattern recognition through successive iterations. The first iteration performs the clustering of the whole sample (load curves from the database) based on specific features associated with the consumption profile, and some clusters of load curves are obtained. The subsequent iterations consider only the medians (patterns) of each group generated and verify the similarity between these medians based on the same statistical tests considered in the first iteration such that some patterns may be collapsed. Thus, at the end of the first phase (after convergence), patterns or types associated with load curves are recognized. The second phase defines the final groups associating each load curve (database) to one of the patterns recognized in the first phase.

Initially, each curve is normalized within the interval [0; 1] dividing the hourly measurements by the peak demand of each. The dimensionless consumption quantified in this way is called power per unit (pu) [16].

Table 1 presents the criteria considered in the similarity analysis performed in each step of the first phase together with the statistical test applied. These criteria were established according to the requirements and indicators practiced in the electric energy distribution sector [10,18–21].

The three features (three stages) presented in Table 1 are applied successively. In the first iteration, the clusters are formed based on similarity between the load curves and the curve with the highest average power consumption (reference curve). After the first iteration, the method assumes that the median of each group keeps its own features and the same tests are successively applied considering the medians (patterns). The existence of a similarity between medians according to the statistical tests implies the union of groups and new medians are obtained.

The LF presented in Table 1 is the ratio between the average and maximum demands of a load curve. The LF is an evaluation index for the rational use of electric power by the consumer [20]. From

Table 1
Statistical tests used in sequence in the process of typing.

Clustering criterion	Statistical test
Distribution of probability of hourly demand	Chi-square for goodness of fit [16]
Selection by correlation level between load curve	Independent two-sample *t*-test significance for Pearson correlation coefficients [17]
Selection by mean consumption or load factor	Independent two-sample *t*-test for significance difference in mean [17]

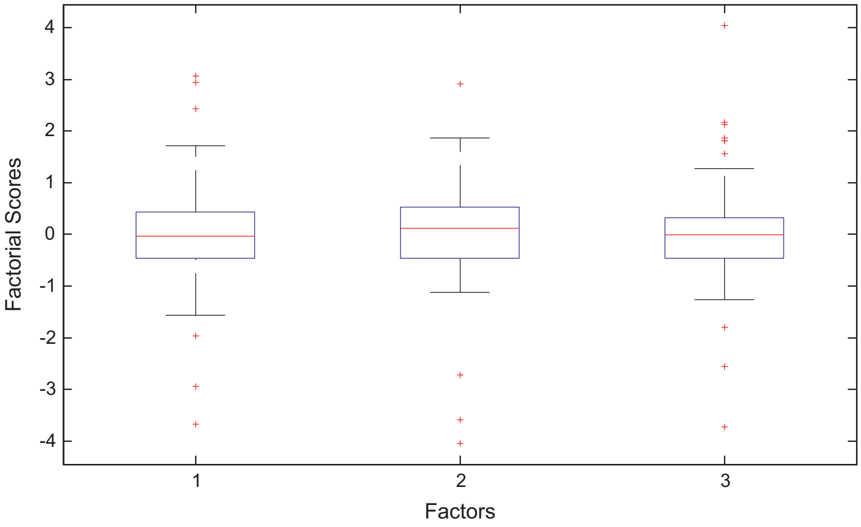

Fig. 2. Distribution of values in each factor (case I).

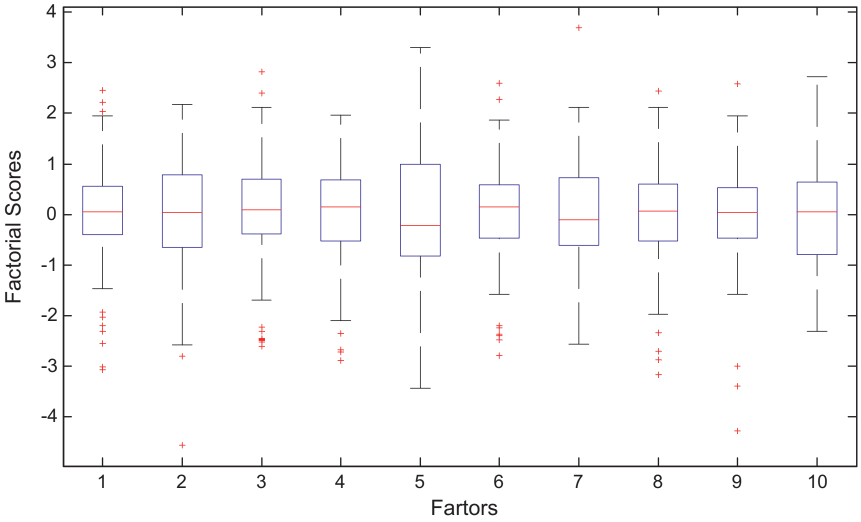

Fig. 3. Distribution of values in each factor (case II).

the viewpoint of the energy distribution system, considering the consumption expectation, the lower the LF, the more rational and economical the energy use will be. From the viewpoint of the equipment, the lower the LF, the more efficient its operation will be.

In the first three stages, the testing of biserial statistical hypotheses between load curves is carried out according to Table 1. There is a fourth stage that comprises a multivariate quantification of dissimilarity between load curves in relation to their dynamic behavior throughout the day. The load curves of each set generated in the third stage are submitted to an additional clustering according to seasonality. This clustering is carried out in two sub stages. In the first, the existence of seasonality is detected through the application of factorial analysis combined with Principal Component Analysis (PCA) of the load curves of each cluster recognized [22]. The analysis of seasonality is performed through the relationship between the hourly consumption of all the curves during the period of 24 h. The data is represented by a nc × 24 matrix where nc is the number of curves presented in each cluster recognized at the end of stage 3. The factorial analysis method applied to this matrix provides the identification of a reduced number (less than 24 and suggested by PCA) of factors that characterize the seasonality of each curve [23]. The second sub-stage comprises the application of a clustering method (subtractive data algorithm [24]) on these factors.

The first phase is repeated successively in order to verify any possible similarity between some of the patterns (median of each group), indicating the need for re-clustering. This first phase is concluded when there is a convergence in the number of patterns. Thus, the number of patterns is a result of the method itself avoiding the need for an initial estimation.

In the second phase of the STCL method (Fig. 1) each load curve of the whole sample is associated to one of the typical curves recognized in the first phase according to the shortest Euclidian distance. The final clusters obtained undergo evaluation. One of the metrics adopted to measure the clustering quality is the Silhouette Index [25]. This index measures the cohesion within and differences between the clusters regardless of the clustering method applied.

Considering N_K load curves (objects) belonging to the K cluster and a total of G clusters ($G \geqslant 2$), the Silhouette Index for each load curve is

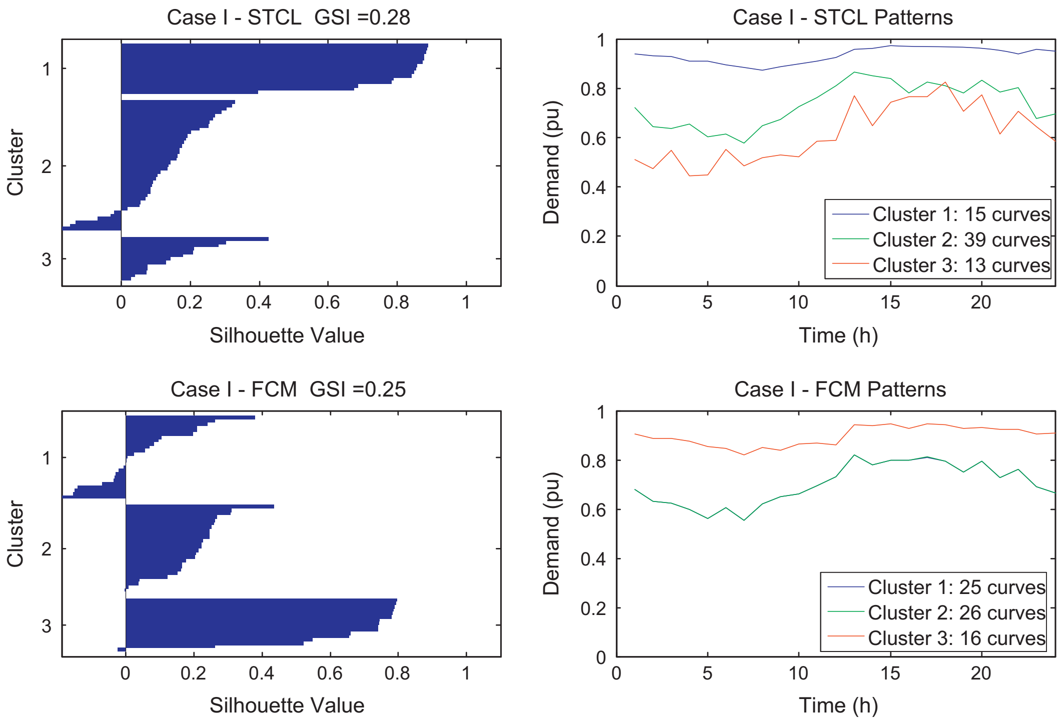

Fig. 4. Silhouette Indices and patterns recognized via the STCL and FCM methods without outlier curves in the sample (case I).

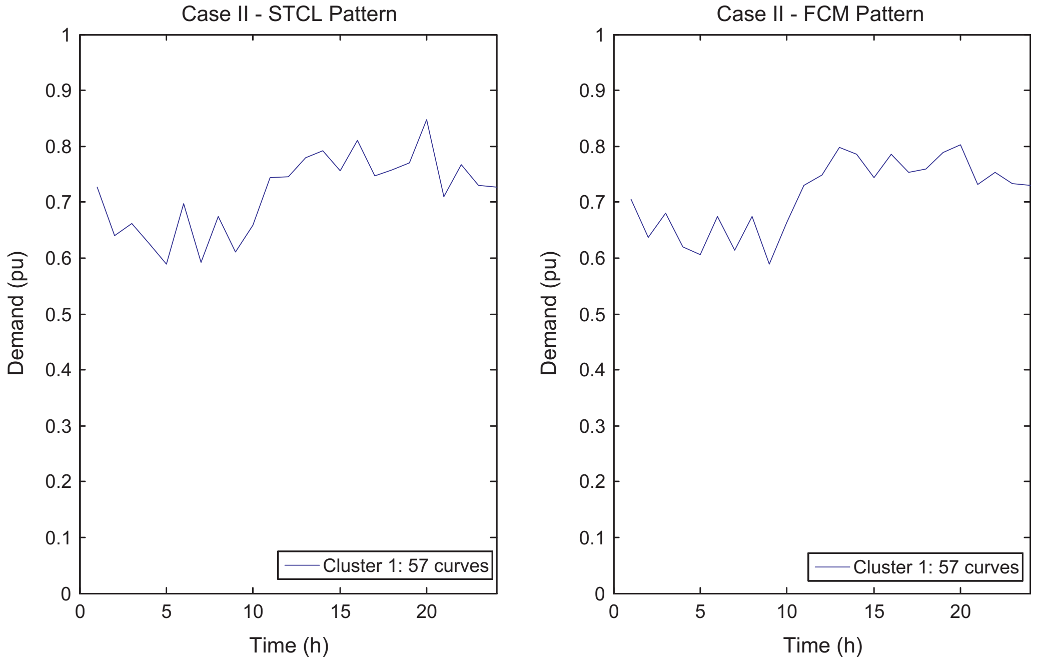

Fig. 5. Patterns recognized via the STCL and FCM methods (case II).

$$S_i^L = \frac{b_i^L - a_i^L}{\max\left\{a_i^L, b_i^L\right\}} \qquad i = 1,\ldots,\sum_{k=1}^{G} N_k \tag{1}$$

where $S_i^L(-1 \leqslant S_i^L \leqslant 1)$ is the Silhouette Index of i curve belonging to the L cluster $(1 \leqslant L \leqslant G)$. a_i^L (Eq. (1)) is the average distance between the i curve and all other load curves belonging to the same cluster. b_i^L (Eq. (1)) is the minimum average distance between the i curve and the load curves belonging to the other clusters.

$$a_i^L = \frac{\sum\limits_{\substack{j=1\\ j\neq i}}^{N_L-1} d_{ij}^{L,L}}{N_L - 1} \qquad i = 1,\ldots,N_L \tag{2}$$

$$b_i^L = \min_{\substack{Z=1\\ Z\neq L}}^{W} \left(\frac{\sum\limits_{j=1}^{N_L} d_{ij}^{L,Z}}{W} \right) \qquad \text{where } W = \sum_{\substack{K=1\\ K\neq L}}^{G} N_K \tag{3}$$

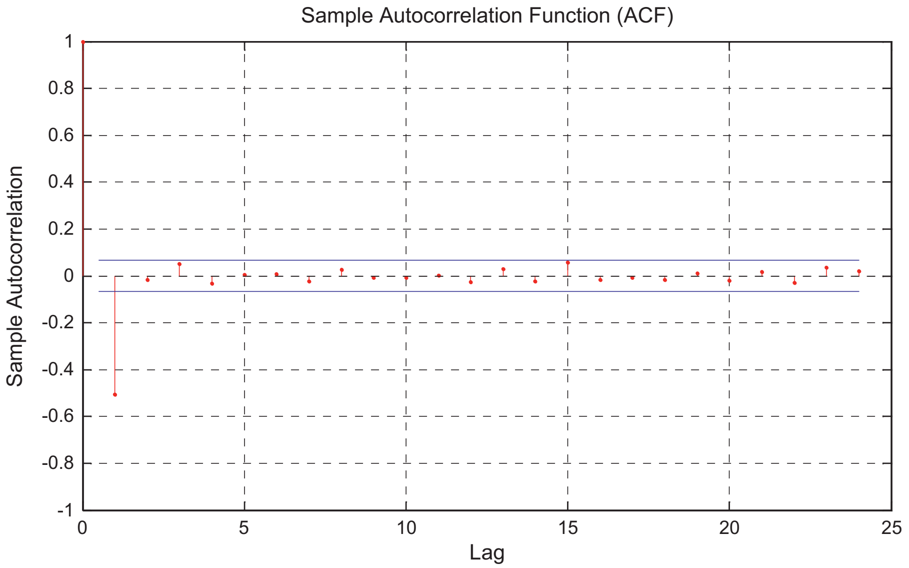

Fig. 6. Autocorrelation values of the first order difference (modal cluster – case I).

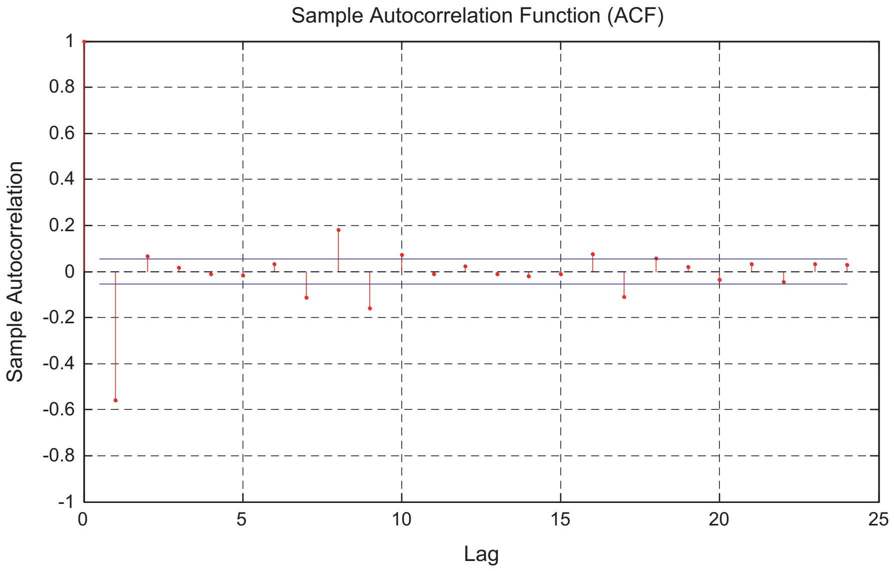

Fig. 7. Autocorrelation values of the first order difference (case II).

$d_{ij}^{L,K}$ is the Euclidian distance between the *i* curve, belonging to the *L* cluster, and the *j* curve belonging to the *K* cluster.

A Silhouette Index close to unity and positive is desired. Negative values imply that there is more homogeneity between the clusters and less internal cohesion. In this work, the mean between the Silhouette Indices [25] obtained for all curves (General Silhouette Index – GSI) was adopted to evaluate the clustering quality.

3. Case study & results

The SCTL method was applied in order to analyze possible changes in the consumption profiles of an energy efficiency program carried out by the Energy Company of Maranhão (CEMAR-Brazil) during the period of November 2008 to July 2009. This program essentially involved the replacement of 5250 old refrigerators for new ones in low-income communities. The sampling process comprised two steps, namely, sampling by clustering and systematic sampling. In the former, the municipalities belonging to the State of Maranhão were classified in clusters according to the similarity between the consumption profiles (load curves). One municipality (center or prototype) was selected to represent each cluster. In the systematic sampling, sampling units (refrigerators) were selected based on the average monthly consumption available in the CEMAR'S records. A sample of eighty load curves (old refrigerators), presenting a high consumption of electric energy (case I, average consumption of 82 kW h) and another sample of 80 load curves after the replacement of the refrigerators (case II, average consumption of 52 kW h) were obtained. In order to reduce the effects of seasonality, all the data are related to working days and the period considered comprises almost the entire summer (the season with the highest energy consumption). The sample size represents an error level of 10% variation in sample means and a confidence level of 95% in the prediction of the population parameter. The International Performance Measurement & Verification Protocol (IPMV) recommends a sampling error of up to ±10% [26].

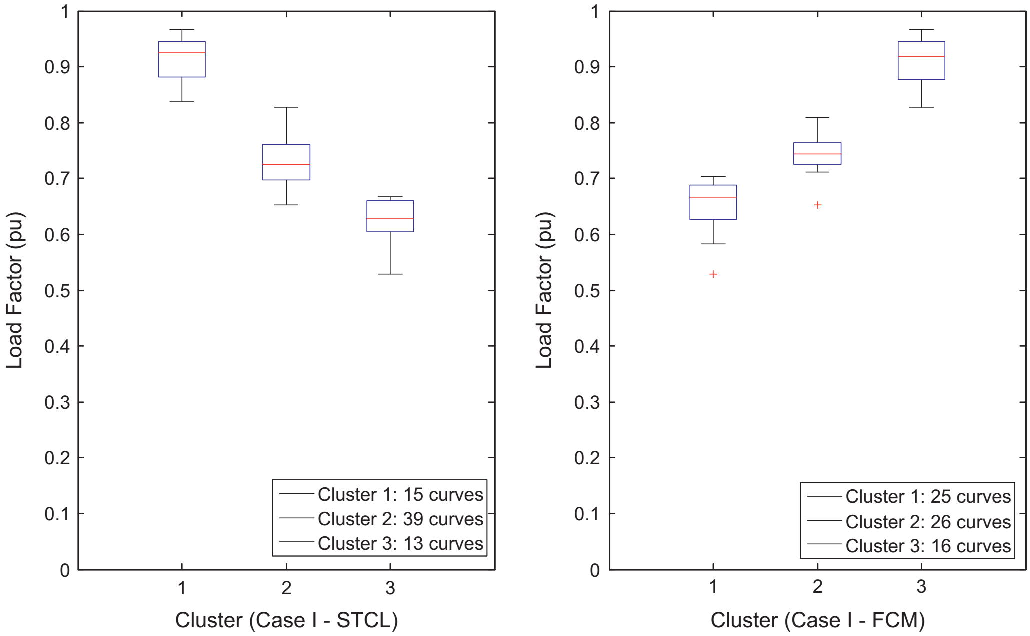

Fig. 8. Distribution of load factor in the clusters recognized (case I).

Initially the data was analyzed to identify and exclude outlier curves. This analysis comprised the application of factorial analysis together with PCA [22,27] (also used in the first phase of STCL method). PCA was used initially to suggest the number of factors that explain 80% of the total variance of the sample (above this level the addition of more factors represented little contribution to explaining the total variance). The PCA provides a reduction in the dimensionality of each load curve (24 points) enabling the selection of factors (less than 24) that can represent the dynamic behavior of each curve. The factorial analysis was then applied to obtain the factorial scores (80 Scores for each factor). Figs. 2 and 3 show the distribution value of factorial scores of the factor for the cases I and II. According to the box-plots, the points identified by the sign "+" indicate outlier curves. Each point may occur on more than one factor (more than one box-plot) associated to the same curve. This analysis enabled the identification of 13 outlier curves in case I and 23 outlier curves in case II. The increase in the number of factors (number of box-plots) in case II is associated to the greater seasonal variation caused by the lower energy demand of the motors of the new refrigerators. This behavior is further corroborated below.

The results obtained using the STCL method were compared with the *Fuzzy C-Means* (FCM), a well-known method belonging to the C-Means families of batch clustering models [28,29], suitable for clustering objects represented by time series [30]. The FCM method requires an initial guess for the number of clusters and, in this case, the same number of clusters provided by the STCL method was considered. The confidence level adopted in the first three stages of the first phase of STCL was 99% in cases I and II.

The application of the STCL method in case I was capable of recognizing the existence of three patterns or demand profiles. On the other hand, two of the three patterns recognized by the FCM were similar attesting the recognition of only two patterns (Fig. 4). The STCL method was capable of recognizing a third pattern of consumption related to 13 curves. This result suggests the ability of STCL to handle a sample of objects with a higher level of heterogeneity (before the replacement of refrigerators). Furthermore, the third pattern represents a profile with lower energy consumption even considering the use of old refrigerators. The quality of clustering obtained with the FCM method was slightly lower according to the Silhouette Index (GSI equal to 0.25 and 0.28 for FCM and STCL methods respectively).

For the sample of load curves after refrigerator replacement (case II), both STCL and FCM recognized the existence of only one cluster and similar patterns (Fig. 5). This shows that the electric energy demand profiles became more similar after the refrigerator replacement, indicating an increase in uniformity among consumers.

Each cluster recognized may be regarded as a sample of possible trajectories that can also be considered as a stochastic process (sample space of possible trajectories [30–32]). From this perspective, the ensemble of load curves of each cluster is a known sample of the underlying stochastic process. In this case, one can check the seasonal variations in each cluster through autocorrelation analysis [31–33]. The existence of two levels of trend in the patterns recognized (specially in relation to case II) suggests, in this case, the application of autocorrelation analysis on the first order differences of the original series in order to mitigate non-stationary effects [31–33]. Fig. 6 presents the correlogram associated to the modal cluster (highest number of load curves) of case I. Only the first value of the autocorrelation (at lag 1) is significant indicating that seasonal variations are due to random factors. In case II (Fig. 7) some autocorrelation values at lags higher than 2 are close to zero and, at the same time, some non-successive autocorrelations are significant, confirming the existence of two more pronounced levels of consumption and the increase in seasonal variations, also confirmed in the preliminary analysis of the curves (factorial analysis). In this case, the increase in the seasonal variations is predictable and consistent because the new refrigerator has better thermal insulation meaning that the motor consumes less power.

An additional analysis comprised the distribution of load factor in the clusters recognized. In case I before optimization (Fig. 8), the STCL method provided clusters whose median load factors is close to the load factor of the respective pattern recognized (0.65, 0.75 and 0.95 pu for the clusters 3, 2, and 1, respectively – Fig. 4). This does not occur with FCM method revealing an inconsistency in the recognition of the patterns (typical curves) in this case. According to STCL method (and also FCM), the distribution of load factors in

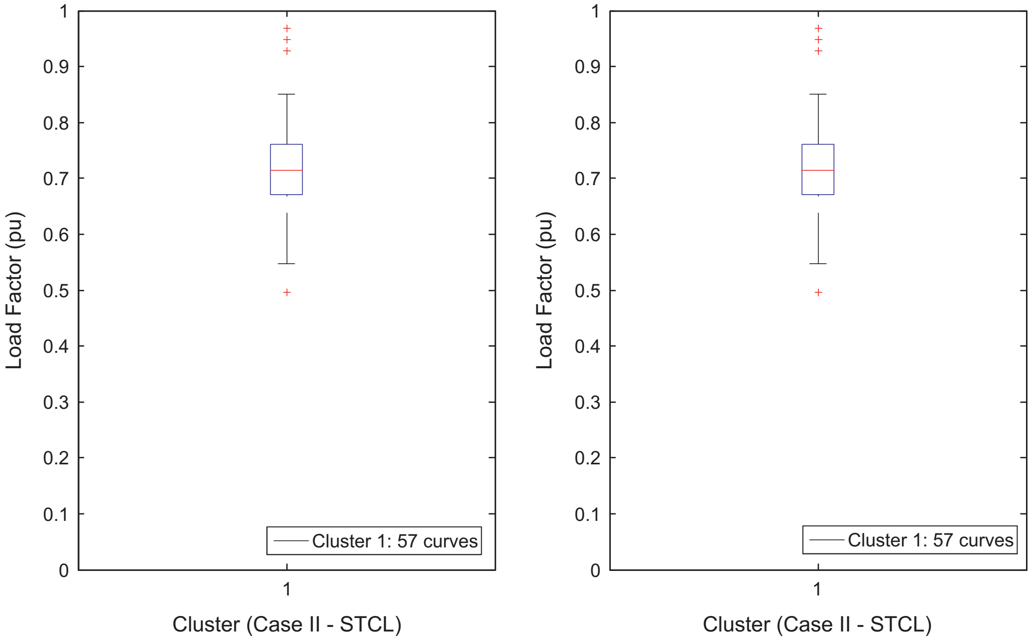

Fig. 9. Distribution of load factor in the clusters recognized (case II).

case II shows lower values (Fig. 9). The reasons are related to the same factors that increase seasonality, i.e. new refrigerators make use of more advanced technology such as better insulation and a motor with a lower energy demand.

4. Conclusion

This study presented a new method of selection, typification, and clustering of load curves (STCL) in which the extraction of features is based on indicators and parameters intrinsic to the electricity sector. The STCL method is suitable for unlabeled data and its adherence to the electric sector makes this a potential tool for applications in this area. The results demonstrate its good performance in recognizing patterns in samples with heterogeneous data (common situation in the electric sector). Unlike C-means models of clustering, the number of clusters is also a result obtained by STCL method.

The case studied looked at an energy efficiency program carried out by the Energy Company of Maranhão (CEMAR-Brazil) which analyzed the impact of replacing 5250 old refrigerators with new ones for low-income consumers. The results obtained by STCL, compared to a well-known method of clustering (*Fuzzy C-Means*, FCM), reveal the viability and potential of the former in recognizing patterns and in generating conclusions coherent with the reality of the electric power sector. This supports the implementation of efficiency actions based on real features within the consumer market and can also support decision-making at management level.

References

[1] Nabeel I, Tawalbeh A. Daily load profile and monthly power peaks evaluation of the urban substation of the capital of Jordan Amman. Int J Electr Power Energy Syst vol. 37, 1, p. 95–102.

[2] Tsekouras GJ, Tsaroucha MA, Tsirekis CD, Salis AD, Dialynas EN, Hatziargyriou ND. A database system for power systems customers and energy efficiency programs. Int J Electr Power Energy Syst 2011;33(6):1220–8.

[3] Jota Patricia RS, Silva Valéria RB, Jota Fábio G. Building load management using cluster and statistical analyses. Int J Electr Power Energy Syst 2011;33(8):1498–505.

[4] Monedero Iñigo, Biscarri Félix, León Carlos, Guerrero Juan I, Biscarri Jesús, Millán Rocío. Detection of frauds and other non-technical losses in a power utility using Pearson coefficient, Bayesian networks and decision trees. Int J Electr Power Energy Syst 2012;34(1):90–9.

[5] Lin JK, Tso SK, Ho HK, Mak CM, Yung KM, Ho YK. Study of climatic effects on peak load and regional similarity of load profiles following disturbances based on data mining. Int J Electr Power Energy Syst 2006;28(3):177–85.

[6] Piatetsky G. Data mining and knowledge discovery 1996 to 2005: overcoming the hype and moving from "universty" to "business" and "analystics". Data Min Knowled Discov 2007;15:99–105.

[7] Gerbec D, Gasperic S, Smon I, Gubina F. A methodology to classify distribution load profiles. Presente dat the IEEE 2002:848–51.

[8] Jain AK, Murty MN, Flynn PJ. Data clustering: a review. ACM Comput Surv 1999;31(3):264–323.

[9] Geminagnani MMF, Oliveira CCB, Tahan CMV. Proposition and comparative analysis of alternative selection and classification of load curve for defining types for tariff studies. Décimo Tercer Encuentro Reginal Iberoamericano de Cigré – XIII ERIAC; 2009. p. 1–6.

[10] Zalewski W. Aplication of fuzzy inference to electric load clustering. New Delhi: IEEE International Conference on Power Systems; 2006. p. 1–5.

[11] Nizar AH, Dong ZY, Zhao JH. Load profiling and data mining techniques in electricity deregulated market. In: Presented at the IEEE power engineering society (PES) general meeting 2006, Montreal, Quebec, Canada; June 2006. p. 1–7.

[12] Han J, Pei J, Yiwen Y. Mining frequent patterns without candidate generation. In: Proceedings ACM-SIGMOD international conference on management of data. ACM Press; 2000. p. 1–12.

[13] Silva Daswin, Yu Xinghuo. A data mining framework for electricity consumption analysis from meter data. IEEE Trans Indust Inf 2011;7(3):399–407.

[14] Gerbec D, Gasperic S, Smon I, Gubina F. Determining the load profiles of consumers based on fuzzy logic and probability neural networks. IEEE Proc Gener Transm Distrib 2004;151(3):395–400.

[15] Zuhaina Zakaria, Lo KL, Hadi Mohamad Sohod. Application of fuzzy clustering to determine electricity consumers' load profiles first international power and energy conference. Putrajaya, Malaysia; 2006. p. 99–103.

[16] Anuar N, Zakaria Z. Cluster validity analysis for electricity load profiling. In: IEEE international conference on power and energy. Kuala Lumpur Malaysia; 2010. p. 35–8.

[17] Cursino dos Santos Arthur Henrique, Werneck Fagá Murilo Tadeu, Moutinho dos Santos Edmilson. The risks of an energy efficiency policy for buildings based solely on the consumption evaluation of final energy. Int J Electr Power Energy Syst 2013;44(1):70–7.

[18] Anssi Seppālā. Statistical distribution of customer load profile. IEEE, CATALOGUE No. 95TH8130;1995. p. 696–701.

[19] Joanicjusz Nazarko, Styczynski Zbigniew A. Application of statistical and neural approaches to the daily load profiles modeling in power distribution systems. IEEE 1999:320–5.

[20] Joseph Janes. Categorical relationships: chi-square. Library Hi Technol 2001;19(3):296–8.

[21] O'Gorman TW. A comparison of an adaptive two-sample test to the *t*-test, rank-sum, and log-rank tests. Commun Statis – Simul Comput 1997;26:1393–411.

[22] Motomasa DAIGO. Factor analysis and pattern decomposition method. SPIE 2005;6043(604317):1–8. 53–65.

[23] Yu Daren, Yu Xiao, Hu Qinghua, Liu Jinfu, Anqi Wu. Dynamic time warping constraint learning for large margin nearest neighbor classification. Inf Sci 2011;181:2787–96.

[24] Chiu S. A cluster estimation method with extension to fuzzy model identification. In: Proceedings of the third IEEE conference on fuzzy systems, vol. 2, Orlando – Florida, USA; 1994. p. 1240–5.

[25] Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math 1987;20:53–65.

[26] International performance measurement & verification protocol – IPMV; 2007.

[27] Hubert M, Vandervieren E. An adjusted boxplot for skewed distributions. Comput Stat Data Anal 2008;52:5186–201.

[28] Bezdek James C, Keller James, Krisnapuram Raghu, PAL Nikhil R. Fuzzy models and algorithms for pattern recognition and image processing. Springer Science+Businees Media, Inc.; 2005.

[29] Bensaid A, Hall LO, Bezdek James C, Clarke LP. Partially supervised clustering for image segmentation. Patt Recogn 1996;29(5):859–87.

[30] Warren Liao T. Clustering of time series data-a survey. Patt Recogn 2005;38(11):1857–74.

[31] Abdel-Aal RE. Modeling and forecasting electric daily peak loads using abductive networks. Int J Electr Power Energy Syst 2006;28(2):133–41.

[32] Saini LM, Soni MK. Artificial neural network-based peak load forecast tingusing conjugate gradient methods. IEEE Trans Power Syst 2002;17:907–12.

[33] Aggarwal Sanjeev Kumar, Saini Lalit Mohan, Ashwani Kumar. Electricity price forecasting in deregulated markets: a review and evaluation. Electr Power Energy Syst 2009;31:13–22.