Electrical Power and Energy Systems 67 (2015) 613–626

Contents lists available at ScienceDirect

Electrical Power and Energy Systems

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

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

Adonias Magdiel Silva Ferreira [a,*], Cristiano Hora de Oliveira Fontes [a], Carlos Arthur Mattos Teixeira Cavalcante [a], Jorge Eduardo Soto Marambio [b]

[a] *Program of Industrial Engineering, Polytechnic School, Federal University of Bahia, Salvador, BA, Brazil*
[b] *Norsul LTD, Norsul LTD, Salvador, BA, Brazil*

ARTICLE INFO

Article history:
Received 4 September 2013
Received in revised form 27 November 2014
Accepted 1 December 2014
Available online 27 December 2014

Keywords:
Clustering
Pattern
Electricity distribution
Multivariate time series

ABSTRACT

This work presents a new method for the clustering and pattern recognition of multivariate time series (CPT-M) based on multivariate statistics. The algorithm comprises four steps that extract essential features of multivariate time series of residential users with emphasis on 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 Electric Company of Alagoas (Brazil) that considers, among others, the analysis of the impact of replacing refrigerators in low-income consumers' homes in several towns located within the state of Alagoas (Brazil). The results were compared with a well-known method of time series clustering already established in the literature, the Fuzzy C-Means (FCM). Unlike C-means models of clustering, the CPT-M method is also capable to obtain directly the number of clusters. The analysis confirmed that the CPT-M method was capable to identify a greater diversity of patterns, showing the potential of this method in better recognition of consumption patterns considering simultaneously the effect of other variables in additional to load curves. This represents an important aspect to the process of decision making in the energy distribution sector.

Introduction

In the current model of regulation of the Brazilian electric sector, programs are often launched by the government to improve the performance of the electric energy distribution sector. These are related to energy efficiency in particular, which in turn is encouraged through tax benefits. One such program comprises the exchange of old refrigerators for new ones in low-income communities. The most common way to evaluate the achievement of goals of energy efficiency is to compare the behavior of the power consumption (load curve) including possible displacement of the peak hours before and after the replacement of refrigerators. In order to improve the quality of the final evaluation of these programs, the load curve must be analyzed together with other time series, increasing the level of knowledge about demand behavior in the electric system [1].

In the management of demand in the electrical system, some variables influence the consumption patterns of electricity [2,3]. Silk and Joutz [4] present a list of factors (price of electricity, price of home appliances, charging, dependence on energy, geographic location, ambient temperature, among others) that affect the electricity demand. On the other hand, the characterization of power consumption in the Brazilian electric sector is based on the charges imposed by the government regulatory agency. In the tariff class of residential consumers, in particular low-income consumers, refrigerators have the highest impact on energy consumption. The ambient temperature in turn has a strong effect on the thermal efficiency of the refrigerator, especially with respect to the use mode (frequency of opening the refrigerator door) which contributes to internal temperature deviations from the set point [5–7]. The thermal efficiency in this case can be quantified through the inverse of the Coefficient of Performance (COP^{-1}) of Carnot, which expresses the ratio between the work for cooling and the heat absorbed from the cold source [8]. Other meteorological variables (such as air temperature, humidity and rainfall) affect the energy consumption in homes as a whole [9] but, for the case study (consumption in refrigerators), just the room temperature should be considered.

* Corresponding author.
E-mail addresses: adoniasmagdiel@ufba.br, adoniasmagdiel@gmail.com (A.M.S. Ferreira), cfontes@ufba.br (C.H. de Oliveira Fontes), arthurtc@ufba.br (C.A.M.T.), norsul@norsulengenharia.com.br (J.E.S. Marambio).

http://dx.doi.org/10.1016/j.ijepes.2014.12.001
0142-0615/

Despite many works on cluster analysis in univariate time series [10,11]), the standard approaches [10] and the feasibility to treat this kind of problem using point-prototype clustering models [12], the pattern recognition in multivariate time series represents a more complex problem (non-point prototyping problem) with intrinsic features [13]. This kind of problem cannot be solved directly using classic models of clustering point-prototype such as Fuzzy C-Means (as can be seen in Section 'Multivariate time series and objects'), requiring special methods in this situation [14–20]. Furthermore, additional challenges must be considered such as the extraction of features from each data/object (set of time series), the similarity metrics related to the domain adopted and the approaches (featured-based or model-based, [12]).

Despite the characterization of energy consumption in the electric sector, there is a lack of works in the pattern recognition using multivariate time series. Chicco [21] present a review about clustering methods (adaptive vector quantization [22], entropy-based, Renyi [23], follow-the-leader [24–29], fuzzy logic [30], fuzzy and ARIMA [31], fuzzy k-means [32,33,22], hierarchical clustering [26,29,22,33], iterative refinement clustering [35], k-means [32,34,26,29,22], min–max neuro-fuzzy [36], multivariate statistics (MANOVA) [37], probabilistic neural network [30,38], self organizing map [32,25,34,26,37,30,39–41], support vector clustering [42] and weighted evidence accumulation clustering [33]) all applied to the univariate case (only load curves). One can divide these methods into two major groups, namely, hierarchical and non hierarchical methods [43–46] and methods based on artificial intelligence [47–49]. A common feature of these methods is that the number of clusters (number of patterns) need to be previously defined. Recent works argue that methods based on artificial intelligence, specifically on C-means models [12], provide better quality in the clustering of load curves [50–56]. All the methods and works cited cope with the electric power consumption separately from other variables resulting in a point-prototype clustering.

Pattern recognition associated to the electric power consumption carried out together with other variables comprises a non point-prototype problem where each object is a set of time series. This work presents a new method of selection, pattern and clustering of multivariate time series (CPT-M) based on a systematic extraction of features from the objects. The case study analyzed comprises an energy efficiency program carried out by the Electric Company of Alagoas (Brazil), that considers, among others, the analysis of the impact of replacing refrigerators in low-income consumers' homes. The Energy Efficiency Program (EEP) was established by the National Agency of Electrical Energy (NAEE) in order to mitigate electricity losses. The main actions of this Program comprise the replacement of low efficiency end-use equipment (refrigerators) with new ones the developing of strategies to raise awareness of local population with regard to the rational and safe use of electricity. The eligible users must meet the following requirements: have a single-phase residential connection, have no irregular power connections, live in the region covered by the Program, have no debts with the electricity utility and have an average consumption in the last three months of more than 59 kWh (specific requirement for the replacement of refrigerators). Furthermore, the refrigerators removed from the customers' homes are recycled and do not return to the consumer market. In this case, the method is able to recognize patterns of energy consumption conjugated to the dynamic profiles of outside and fridge temperatures also providing a consistent way to evaluate the efficiency of the program. The proposed method incorporates multiple criteria in the clustering and pattern recognition in load curves unlike traditional approaches that essentially use the criterion of distance between load curves [34]. Section 'The CPT-M method' presents a new method for the clustering and pattern recognition of multivariate time series (CPT-M) and the evaluation metrics adopted [57]. Section 'Case study and results' presents the case study and results obtained from a new version of Fuzzy C-Means – FCM (with algorithm adapted to the non point-prototype problem [58]) and CPT-M methods showing the ability and superiority of the latter to the problem analyzed.

Multivariate time series and objects

In Data Mining Theory, Clustering is the task of grouping data (or objects) into clusters according to the principles of homogeneity (data or objects belong to the same cluster should be as similar as possible) and heterogeneity (data or objects belong to different clusters should be as different as possible) [59,60]. The clustering problem comprises unsupervised learning as there are no pre-labeled objects (there is no previous information to distinguish the objects from each other) [17,45,60].

A sample with n objects can be represented by the set $X = \{x_1, x_2, \ldots, x_n\}$. If each object $x_i (i = 1, \ldots, n)$ is a feature vector in the space $\Re^p$ (p is the dimensionality of data set) there is the problem of point-prototype clustering [12].

A time series is a series of observations (measurements) made sequentially through time and associated to a specific process variable [60]. Two kinds of objects must be previously considered and represent different problems of clustering and pattern recognition of time series, namely, the Univariate Time Series (UTS) and Multivariate Time Series (MTS) [15,61,62]. Considering a general series of observations over time $z_i(t) (i = 1, \ldots, k;\ t = 1, \ldots, m)$ where k is the number of variables (number of sensors), m is the number of observations and i indexes the measurements made at each time instant, a UTS object comprises the case in which $k = 1$. Otherwise ($k \geqslant 2$) there is a MTS object. Univariate time series has been broadly explored and is often regarded as a point-prototype clustering problem in multidimensional space ($\Re^m$). Traditional metrics of similarity (Euclidian distance) and traditional clustering methods applied to clustering static data (such as Fuzzy C-Means and k-Means) can also be used in this case (raw-data-based approach [63]). On the other hand, MTS objects are common in various areas of knowledge. This approach is mandatory when there is need to consider more than one variable and all of them in an integrated way. An MTS should be treated as a whole and may not be transformed into one long univariate time series [61]. The clustering of MTS comprises a non-point prototyping problem and traditional metrics of similarity, appropriate for the univariate case, cannot be applied in this case. Additional challenges should be considered in the clustering of MTS objects such as feature extraction, similarity metrics and the selection of appropriate variables (reduction in the size of the problem) [14,16,13,17,15,58]. Despite feature extraction and similarity metrics, and the fact that traditional metrics cannot be used, specific alternatives based on Principal Component Analysis (PCA) and Wavelets Transform have been proposed [15,64]. The selection of appropriate variables can be performed either through the use of specific techniques (such as PCA) or also by an analysis of system and profiles of the time series.

A MTS object can be represented by the following $m \times k$ matrix:

$$Z_i = \begin{bmatrix} z_{i1}(1) & \cdots & z_{ik}(1) \\ \vdots & \ddots & \vdots \\ z_{i1}(m) & \cdots & z_{ik}(m) \end{bmatrix}$$

where Z_i is the object, $z_{ij}(t)$ is the measurement of variable j ($j = 1, \ldots, k$) at time instant t ($t = 1, \ldots, m$) in the object Z_i ($i = 1, \ldots, n$ objects). The column j contains the time series related to the variable j.