XXIV CONGRESSO
SOCIEDADE PORTUGUESA DE ESTATÍSTICA

ATAS

Estatística: Desafios Transversais às Ciências com Dados

COMISSÃO EDITORIAL
PAULA MILHEIRO
ANTÓNIO PACHECO
BRUNO DE SOUSA
ISABEL FRAGA ALVES
ISABEL PEREIRA
MARIA JOÃO POLIDORO
SANDRA RAMOS

SPE Sociedade Portuguesa de Estatística
P.PORTO
ESCOLA SUPERIOR DE TECNOLOGIA E GESTÃO
isep Instituto Superior de Engenharia do Porto

Editores: Paula Milheiro, António Pacheco, Bruno de Sousa, Isabel Fraga Alves, Isabel Pereira, Maria João Polidoro e Sandra Ramos

Título: Estatística: Desafios Transversais às Ciências com Dados
Atas do XXIV Congresso da Sociedade Portuguesa de Estatística

Editora: Sociedade Portuguesa de Estatística

Conceção gráfica da capa: Ana Ferrás (ESTG - IPP)

ISBN: 978-972-8890-47-6

Combining various dissimilarity measures for clustering electricity market prices

Margarida G. M. S. Cardoso
Instituto Universitário de Lisboa (ISCTE-IUL), Business Research Unit (BRU-IUL), *margarida.cardoso@iscte-iul.pt*

Ana Martins
Instituto Superior de Engenharia de Lisboa (ISEL), *ana.martins@isel.pt*

João Lagarto
Instituto Superior de Engenharia de Lisboa (ISEL), INESC-ID, *joao.lagarto@isel.pt*

Keywords: Electricity markets; Time series; Clustering validation

Abstract: The analysis of electricity markets of the European countries aims to better understand their degree of integration, which is relevant for the development of an internal market of electricity in the European Union.
This study resorts to clustering of time series of hourly prices of electricity (in €/MWh) observed in the day-ahead market in 2018. The proposed approach relies on the combination of different dissimilarity measures which can capture differences in time series trends, (prices) values, cyclical behaviors and autocorrelation patterns. The results obtained, enable to provide some insights on the role of the different dissimilarity measures in the clustering process. Furthermore, they provide a clustering solution with coherent substantive interpretation and, interestingly, one that reveals the natural patterns of geographic proximity.

1 Introduction

The development of an internal market of electricity has long been a goal of the European Union, since it enables European citizens and businesses to choose their supplier, creates new business opportunities and enhances cross-border trade with the purpose of ensuring efficiency gains and competitive prices, and contributes to security of supply and sustainability.

In electricity markets, power producers offer their electricity production at a given price. Producers' offers are ordered by economic merit order, i.e., from the lowest to the highest price, and this gives rise to the supply curve. Buyers (retailers and big consumers) bid for the electricity they want to acquire at a given price. Buyers' bids are also ordered by economic merit order, which in this case is from the highest to the lowest price, thus generating the demand curve. The market price is defined as the price for which the supplied quantity is equal to the bought quantity, i.e., market price and quantity are set at the intersection of the supply and demand curves. Shifts in demand and in supply induce different hourly prices, i.e., price variability during the day.

In order to better understand the degree of integration of the electricity markets of different European countries, we cluster time series regarding each country's day-ahead electricity market prices.

The electricity markets under study are the MIBEL, the Italian, the Nord Pool, the French and the German markets. The MIBEL is the Iberian electricity market and has two price regions: Portugal and Spain, MI_PT and MI_ES, respectively. Italy is divided into six price regions: IT_CNOR, IT_CSUD, IT_NORD, IT_SARD, IT_SICI and IT_SUD. The Nord Pool market incorporates seven countries: Sweden (with four areas Np_SE1, Np_SE2, Np_SE3 and Np_SE4), Denmark (with two areas Np_DK1 and Np_DK2), Norway (with six areas Np_Oslo, Np_Kr.sand, Np_Bergen, Np_Molde, Np_Tr.heim, Np_Tromsø), Latvia, Lithuania, Finland and Estonia.

Clustering base data are hourly prices of electricity (in €/MWh) for 26 regions of Europe, observed in the day-ahead market in 2018.

This paper is organized as follows. Section 2 presents a brief literature review. In Section 3 the methodological approach is introduced. Section 4 presents the experiments results and the selected clustering solution. Finally, in Section 5, the main conclusions of this work are presented.

2 Literature review

Clustering methods aim to organize a data set into well separated groups, where similar items are within the same group and dissimilar items are in different groups. Since time series data are common in diverse scientific domains, the clustering of time series as been a topic of interest in the literature (e.g.[1]). Several alternative approaches have been proposed for grouping time series, namely concerning the clustering method and the adopted proximity measure between two time series - e.g.[1],[2]. The choice of a dissimilarity or a distance measure is a critical issue in clustering time series and it can be defined by considering the raw time series data, some features vector extrated from data or by comparing the parameters of underlying time series models [3]. Many works have applied clustering methods to extract useful information from the electricity price time series. In [4], different clustering algorithms were applied, in particular Ward hierarchical method and a Self-Organizing Map, to obtain diverse daily profiles of consumption, wind generation and electricity spot prices. These profiles were then used to simulate residential demand response programs and small-scale distributed energy storage systems.
Cluster analysis has also been used as a way of pre-processing the input data for the forecasting of the demand or of the electricity price. The goal is to identify homogeneous groups which later can be used to improve the forecasts and eventually to detect outliers (e.g.[5],[6],[7]).

3 Methodology

The proposed approach relies on the use of different distance measures between time series that will enable to capture diverse aspects of the differences between them. Several experiments are conducted to combine these distances. A clustering method able to deal with the integration of the various dissimilarities measures is presented and a process of evaluating the alternative solutions obtained is addressed. The analysis regards hourly day-ahead electricity prices' time series data referring in 26 European electricity market zones.

3.1 Distance measures

Alternative distance measures between two time series x_t and y_t, $(t = 1,...,T)$ provide different insights regarding the differences between them, which can be combined to provide a better clustering solution. Namely, we consider the following distance measures: Euclidean (d_{Euclid}), a Pearson correlation based measure ($d_{Pearson}$), a Periodogram based measure (d_{Period}) and an Autocorrelation based measure ($d_{Autocorr}$).
The Euclidean distance, (d_{Euclid}), yields the sum of Euclidean distances corresponding to each pair (x_t,y_t) which captures differences in scale.
The Pearson correlation based measure takes into account linear increasing and decreasing trends over time. In this work we resort to a measure suggested in [9]

$$d_{Pearson} = \sqrt{\frac{1 - r_{x_t,y_t}}{2}}, \tag{1}$$

where r_{x_t,y_t} represents the Pearson correlation.
Let $P_x(w_j)$ be the periodogram of time series x_t at frequencies $w_j = 2\pi j/n, j = 1,...,[n/2]$ in the range 0 to π, being $[n/2]$ the largest

integer less or equal to $n/2$,

$$P_x\left(w_j\right) = \left(\frac{1}{n}\right)\left|\sum_{t=1}^{T} x_t e^{-itw_j}\right|^2. \tag{2}$$

The Periodogram based measure [10] considers the Euclidean distances between the Periodograms $P_x\left(w_j\right)$ and $P_y\left(w_j\right)$ of time series x_t and y_t, respectively. It expresses the contribution of the various frequencies or cyclical components to the variability of the series,

$$d_{Period} = \left(\sum_{j=1}^{\left[\frac{T}{2}\right]}\left(P_x\left(w_j\right) - P_y\left(w_j\right)\right)^2\right)^{\frac{1}{2}}. \tag{3}$$

Finally, the Autocorrelation based distance [3] calculates Euclidean distances between autocorrelation structures, comparing the series in terms of their dependence on past observations

$$d_{Autocorr} = \left(\sum_{l=1}^{L}\left(r_l\left(x_t\right) - r_l\left(y_t\right)\right)^2\right)^{\frac{1}{2}}, \tag{4}$$

where $r_l\left(x_t\right)$ and $r_l\left(y_t\right)$ represent the estimated autocorrelations of lag l of $\left(x_t\right)$ and $\left(y_t\right)$, respectively. The four referred distances are implemented in the R package "TSclust" [3]. In order to combine all the distances, each one is firstly normalized using a min-max transformation,

$$norm(d(x_{t_i},x_{t_j})) = \frac{d(x_{t_i},x_{t_j}) - min\{d(x_t,x_t)\}}{max\{d(x_t,x_t)\} - min\{d(x_t,x_t)\}}. \tag{5}$$

Then, a convex combination of the four (normalized) distances is considered in the clustering process, aiming to incorporate the different perspectives that each distance measure provides to the clustering solution.

3.2 Clustering algorithm

The K-Medoids algorithm is adopted for clustering (we use R package "cluster"). It aims at the minimization of the distance of objects belonging to a cluster from the cluster's medoid, for all clusters. It generalizes K-Means using arbitrary-defined distance measures and it is somewhat more flexible in terms of cluster shapes and more robust to outliers and noise than K-Means. In what concerns time-series clustering, the fact that a medoid (a member of the data set) is considered, overcomes the need to determine a centroid (based on an averaging of different series) which can be a problematic issue [8].

3.3 Number of clusters

To determine the number of clusters (K) we resort to several cohesion-separation measures - namely Average Silhouette (Silh) [8], Calinski and Harabasz (CH) [11] and Dunn modified index (Dunn2) [12] (implemented in the "fpc" R package [13]). We also consider the relative improvement or rate of change in within clusters' variation (Winprov), i.e., the total distances between each observation and the corresponding medoid, between two successive solutions (with $k-1$ and k clusters). The first three indices present a variant of a between-within clusters distance and the fourth considers within clusters' distances only.
A summated indicator of all indices is considered. First, each index, ind_k, is transformed by a max-min function, i.e. we use $1 - norm(ind_k)$, with $norm(ind_k)$ defined as in (5). Note that, after transformation, all indices can be viewed as preference values and the lower their values, the better the clustering solution they correspond to. Then, the sum of all the (transformed) indices values referring to the candidates numbers of clusters is considered for selecting the number of clusters. The results obtained are finally evaluated from a substantive point of view.

3.4 Experiments

We consider a convex combination of Euclidean distance, Pearson correlation based distance, Periodogram based distance and Autocorrelation based distance using corresponding weights $\underline{w} = (w_1, w_2, w_3, w_4)$. Several experiments are considered allocating different weights to the alternative distance measures used in the clustering algorithm. The experiments can be described as follows:

1. Calculate and normalize the distances between the time series and define a convex combination of the same distances, using $\underline{w}$;
2. Run K-Medoids for $k = 2,...,10$, using the combined distance;
3. Calculate clustering evaluation indices and obtain the summated indicator of clustering quality, based on the rescaled clustering evaluation indices (for $k = 2,...,10$);
4. Decide on K based on 3.;
5. Graphically represent the obtained solution, interpret it, compare it to other solutions and resort to substantive insights to decide on the "best" solution.

In order to obtain a graphic representation of each clustering solution we use classical Mutidimensional Scaling. A goodness of fit measure (GOF) enables to evaluate the quality of the MDS maps

$$GOF = \frac{\sum_{i=1}^{p} \lambda_i}{\sum_{i=1}^{n-1} |\lambda_i|}, \tag{6}$$

where λ_i are the eigenvalues of $B = X^T X$ and $X_{n \times p}$ represents the p coordinates of the n points to be represented. As in this case we use two dimensions, $p = 2$, in the numerator we have the two largest eigenvalues. The GOF varies between 0 and 1 and the higher the GOF the better the MDS map ([14]).

4 Results obtained

We consider the hourly day-ahead market prices (in €/MWh) observed in 26 european electricity markets zones in 2018. Data were obtained through the respective market operators' website. Due to different winter and summer times in some countries, the raw data have one missing hour, that was fulfilled with the average price of the two nearest hour prices, and a redundant hour data, that was removed.
Several experiments allocating different weights/relevance to the alternative distance measures in the clustering algorithm are reported in Table 1. For each solution some insights are provided.

Table 1: Some results of experiments conducted where $\underline{w} = (w_1,w_2,w_3,w_4)$ stands for weights that correspond to Euclidean distance (1), Pearson correlation based distance (2), Periodograms based distance (3) and Autocorrelation based distance (4) in a convex combination of distances.

$\underline{w}$	K	Comments
(1,0,0,0)	2	Italy and MIBEL markets are gathered into one cluster; the remaining regions are clustered together.
(0,1,0,0)	2	Nord Pool is separated from other regions.
(0,0,1,0)	4	IT_SICI constitutes a single cluster; NP_LV and NP_LT constitute a clearly separated cluster.
(0,0,0,1)	3	IT_SICI is included in a six regions cluster.
$(0,1/3,1/3,1/3)$	6	See "The selected clustering solution".
$(1/3,0,1/3,1/3)$	9	The exclusion of Pearson based distance brings a relevant increase to the solutions' entropy.
$(1/3,1/3,0,1/3)$	6	See "The selected clustering solution".
$(1/3,1/3,1/3,0)$	6	The exclusion of Autocorrelation based distance increases the isolation of the single cluster IT_SICI.
$(1/4,1/4,1/4,1/4)$	6	See "The selected clustering solution".

4.1 The selected clustering solution - The experiment with all distances equally weighted

After conducting the proposed experiments, we arrived at a solution that considers all distances equally weighted, in the convex combination distance used in K-Medoids, and yields 6 clusters - see the indicator based on the sum of the normalized validation indices illustrated in Table 2. To adopt this solution we firstly take into account a consistency criterion: this solution exhibits perfect consistency with solutions (0,1/3,1/3,1/3) - exclusion of Euclidean distance only - and (1/3,1/3,0,1/3) - exclusion of Peridogram based distance only (see index of agreement values in Table 3 and Table 4). In addition to exhibiting consistency, this solution also proves to be a good solution from the point of view of the intra-clusters and inter-clusters variation relationship: the results referring to each clustering validation index are presented in Figure 1. In particular, the Silhouette index exhibits a value over 0.5 which indicates a reasonable partitioning of data (values less than 0.2 means that the data do not exhibit cluster structure) [8]. Finally, this partition of European regions is coherent with their geography - Figure 2 and Figure 3 - and clusters are compatible with domain knowledge. The representants of clusters - the medoids - are illustrated in Figure 4, using the corresponding cronograms.

Table 2: Summated indicator for clustering evaluation for alternative numbers of clusters.

K - number of clusters	2	3	4	5	6	7	8	9	10
Summated indicator	2.26	2.04	2.89	1.76	0.87	2.03	1.85	1.39	1.85

4.2 The clusters

The solution obtained, besides capturing the natural geographic relationship which is intrinsically related with the price regions, also

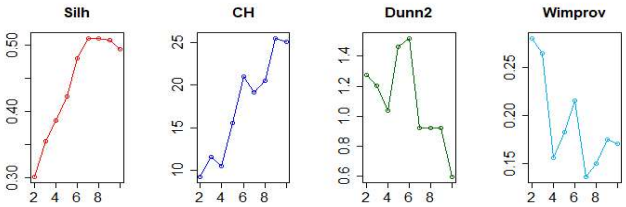

Figure 1: Clustering validation indices for K-Medoids solutions (K = 2,...10) derived from the convex combination of all distances uniformly weighted.

Table 3: Adjusted Rand between clustering solutions (I).

	(0,1,0,0)	(0,0,1,0)	(0,0,0,1)	(0,1/3,1/3,1/3)
(1,0,0,0)	0.704	0.049	0.047	0.269
(0,1,0,0)	1	0.148	0.186	0.348
(0,0,1,0)		1	0.636	0.478
(0,0,0,1)			1	0.626

provides substantive coherent interpretation - see map in Figure 3 and clusters' medoids in Figure 4.

Cluster C1: The cluster that joins mainland Italy and France (with medoid IT_CSUD) was somewhat unexpected. If, on the one hand, France has its electricity grid interconnected with Italy, on the other hand, it is also interconnected with Spain. In fact, after checking all inter-medoids distances between C1 and C6 (including Spain), we concluded their separation is weak and mainly due to Pearson and Autocorrelation based distances.

Cluster C2: A Danish region Np_DK1, is the cluster medoid. Although Denmark is known to have a lot of wind (renewable) energy, it also uses coal and natural gas to produce electricity, which also occurs in Germany. Sweden's NP_SE4 is very close to Denmark.

Table 4: Adjusted Rand between clustering solutions (II).

	(1/3,0,1/3,1/3)	(1/3,1/3,0,1/3)	(1/3,1/3,1/3,0)	(1/4,1/4,1/4,1/4)
(1,0,0,0)	0.206	0.269	0.301	0.269
(0,1,0,0)	0.209	0.348	0.396	0.348
(0,0,1,0)	0.342	0.478	0.570	0.478
(0,0,0,1)	0.410	0.626	0.555	0.626
(0,1/3,1/3,1/3)	0.741	1.000	0.887	1.000
(1/3,0,1/3,1/3)	1	0.741	0.625	0.741
(1/3,1/3,0,1/3)		1	0.887	1.000
(1/3,1/3,1/3,0)			1	0.887

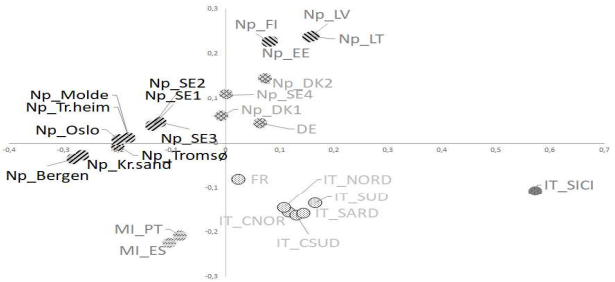

Figure 2: MDS map of equally weighted combination of distances (between price regions) with clusters.

Therefore, to meet all demand of NP_SE4 area, it is necessary to use Danish production which makes prices in NP_SE4 area similar to prices in Denmark and Germany. Interestingly, all distance measures tend to put together the NP_SE4 area and Denmark. The small inter-medoids differences between this cluster and C3 (including other Swedish regions) are due to the autocorrelation based distance.

Cluster C3: This cluster includes Sweden (with the exception of NP_SE4) and Norway geographically close countries using low cost production technologies (renewable and nuclear). Its medoid

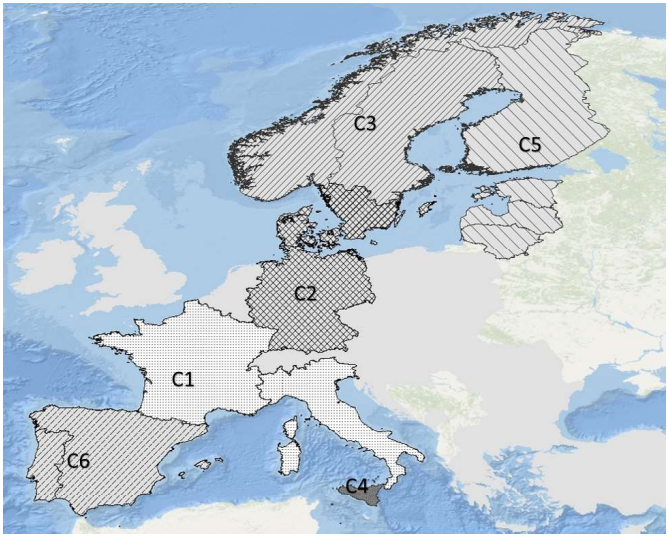

Figure 3: Map of final clusters with medoids.

(Np_Tr.heim) clearly distinguishes itself from C5 medoid (referring to a geographically adjacent cluster) due to large autocorrelation based distance.

Cluster C4: Sicily appears isolated since its interconnection to continental Italy is often limited by the capacity of power lines (an issue already referred). In fact, according to the distances computed it exhibits a large Periodogram based distance to the medoid of C1 cluster (including Italy). It also exhibits maximum Pearson based distances to C5 (emphasizing differences in tendency) and generally differs from all remaining clusters considering all distances.

Cluster C5: Finland and the Baltic countries are geographically close and use fossil technologies (coal and natural gas), although Finland has a lot of nuclear. This clusters' medoid is Np_ LV. Besides differences to C3 medoid already mentioned, it shows moderate

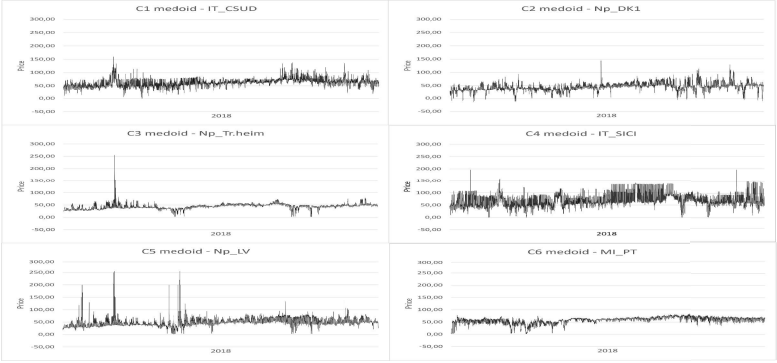

Figure 4: Cronograms of final clusters' medoids.

autocorrelation and Pearson based distances to C1 medoid.
Cluster C6: The grouping of Portugal (the medoid) and Spain is quite natural: geography, interconnection capacity that accommodates flows for most hours and electricity production from the similar technologies - major renewable production (hydro, wind and solar in the case of Spain) and fossils (coal and natural gas). In 2018, in almost 95% of the hours, the prices were the same in Portugal and Spain. The Portuguese time series clearly differs from C3 medoid (due to large Pearson based distance) and from C5 medoid (due to large autocorrelations based distance).

5 Conclusions

The use of K-Medoids algorithm for clustering allowed to overcome the limitations of K-Means that resorts to averages of time series to build centroids and uses squared Euclidean distances. In K-Medoids the representatives of clusters are members of the same clusters and

it enabled exploring the contributions of alternative and combined distance measures. We found that the alternative distance measures played complementary roles by emphasizing differences in trends (Pearson), differences in prices' values (Euclidean), differences in cyclical behaviours (Periodogram) and differences in autocorrelation patterns (Autocorrelation). The combination of all these aspects, ultimately, allowed us to obtain some insights on the sensitivity of the clustering procedure to the role of the distance measures considered and to obtain a good clustering solution. Interestingly, using all distances combined, the clusters obtained reflect not only different prices patterns, but also geographic proximity, which may underline these same patterns. Future work intend to further explore on the tuning of the weights for combining distances, on the decision process concerning the selection of the number of clusters resorting to multiple criteria and on the profiling of clusters obtained using external variables.

Acknowledgements

This work was supported by Fundação para a Ciência e a Tecnologia, grants UIDB /00315/2020 and UIDB/50021/2020.

References

[1] Aghabozorgi, S., Shirkhorshidi, A. S., Wah, T. Y. (2015). Time-series clustering–a decade review. *Information Systems*, vol. 53, pp. 16–38.

[2] Maharaj, E. A., D'Urso, P., Caiado, J. (2019), *Time series clustering and classification*, CRC Press.

[3] Montero, P., Vilar, J. (2014). TSclust: An R package for time series clustering. *Journal of Statistical Software*, 62(1), pp. 1–43.

[4] Miguel, P., Gonçalves, P., Neves, L. , Martins, A. G. (2016). Using clustering techniques to provide simulation scenarios for the smart grid. *Sustainable Cities and Society* 26, pp. 447–455.

[5] Martínez-Álvarez, F., Troncoso, A., Asencio-Cortés, G., Riquelme, J. (2015). A survey on data mining techniques applied to electricity-related time series forecasting. *Energies*, vol.8, 11, 13162–13193.

[6] Jin, C. H., Pok, G., Lee, Y., Park, H. W., Kim, K. D., Yun, U., Ryu, K. H., (2015). A SOM clustering pattern sequence-based next symbol prediction method for day-ahead direct electricity load and price forecasting. *Energy Convers. Manag.*, vol. 90, pp. 84–92.

[7] Panapakidis, I. P., Dagoumas, A. S. (2016). Day-ahead electricity price forecasting via the application of artificial neural network based models, *Applied Energy*, 172, pp. 132–151.

[8] Kaufman, L., Rousseeuw, P. J. (2009). *Finding groups in data: an introduction to cluster analysis.* John Wiley & Sons.

[9] Rodrigues, P. (2008). Hierarchical clustering of time-series data streams.*IEEE Transactions on Knowledge and Data Engineer*, vol. 20, 5, pp. 1–13.

[10] Caiado, J., Crato, N., Peña, D. (2006). A periodogram-based metric for time series classification. *Computational Statistics & Data Analysis*, vol. 50, 10, pp. 2668–2684.

[11] Caliński, T., Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, vol.3, 1, pp. 1–27.

[12] Bezdek, J. C., Pal, N. R. (1998). Some new indexes of cluster validity. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 28, 3, pp. 301–315.

[13] Hennig, C. (2015). Package fpc. URL: http://cran. r-project. org/web/packages/fpc/fpc. pdf (available 08.07. 2017).

[14] Cox, T., Cox, M. (2001). *Multidimensional Scaling*, 2nd Ed., Chapman & Hall/CRC.

[15] *GME, Annual Report 2018.* Available Italian Electricity Market Operator web site: http://www.mercatoelettrico.org/En/MenuBiblioteca/documenti/20190731_GME_RELAZIONE_ANNUALE_EN.pdf.