Clustering of electricity price: an application to the Italian electricity market

Mahmood Hosseini Imani
Department of Energy
Politecnico di Torino
Torino, Italy
mahmood.hosseiniimani@polito.it

Mansour Taheri Andani
Department of engineering
Texas state university
Texas,Usa
mansour.taheriandani@gmail.com

Hamid Taheri Andani

Isfahan university of technology
Isfahan,Iran
Hamid.taheriandani80@gmail.com

***Abstract*— Analyzing the electricity price plays a vital role in market players in deregulated electricity markets. In this regard, proper clustering methods are more beneficial. In this paper, a comparative study of three major clustering algorithms, including K-Means, Fuzzy C-Means, and Hierarchical algorithm on the Italian National Single Price, has been carried out. Moreover, the impact of various parameters of the algorithms on the clustering results has been analyzed. The performance of the clustering methods has been compared through several clustering validity indexes, including the Silhouette index, Calinski-Harabasz index, and Davies-Bouldin indicator. Two distinct patterns consisting of working days and weekends (or holidays) are observable in the particular dataset.**

Keywords*— *Clustering methods, Electricity price, Fuzzy C-Means, Hierarchical algorithm, K-Means, PUN

I. Introduction

After deregulation in the electricity market, electricity prices have become more volatile as a result of competition among market players. To overcome the issue, the precise electricity price prediction models are obtaining major importance.

Different parameters including historical data of electricity price, load consumption and generation of renewable energy sources are affecting the performance of the prediction model [1]–[3]. Moreover, extracting more useful information from historical price data and utilizing them as an input of the forecasting model is more advantageous. To this end, there are different clustering methods to extract useful information from historical price data.

Clustering is a popular unsupervised machine learning method for grouping data points from a large data set into a class according to their characteristics [4]. Clustering the data has plentiful applications in the electricity market such as load consumption clustering, electricity price forecasting, implementation of demand response programs (DRPs) [5], [6].

Numerous clustering algorithms including K-Means [7], Fuzzy C-Means (FCM) [8], [9], Hierarchical method [10], [11] , Self-organizing map [12], [13], Optimum-path forest [14] and Gaussian mixture model (GMM) [15] have been introduced in the power market literature.

Ref. [10] used hierarchical clustering for identifying clusters of similar profiles in UK. The information of the clustering results will be useful for scheduling energy storage system operation. In [7] the authors applied K-Means method to cluster Spanish electricity price data. By clustering the electricity price, all the data are separated on 6 groups. These labeled days will be utilized to predict the day-ahead price in future work. Ref [16] used different clustering methods to cluster daily profile of electricity price in Germany. Then the centers of each cluster are used in modeling time-varying operations in complex energy systems optimization problems as the representative days instead of an entire year of 365 days.

In order to evaluate the performance of the clustering algorithms, several Cluster Validity Indexes (CVIs) considering a range of parameters and conditions are proposed in the literature [17]. The purposes of using CVIs are: comparing the performance of various clustering models [18], defining adequate numbers of clusters [19] and analyzing the impact of the method's parameter on the clustering outcomes [20].

In this paper, the clustering of Italian National Single Price (PUN) according to their daily price patterns are extensively explored. The main focus is to find distinct price patterns which, afterward, can be utilized to improve the performance of the future applications (i.e., electricity price forecasting, design suitable tariffs, and DRPs). In the case of electricity price prediction, the input value, predictors, play a key role to improve the forecasting performance. The more accurate predictors, the more prediction performance. Splitting the various patterns of electricity price in each day by labeling them and using the labels to train historical data can improve the price prediction.

To this end, three major clustering techniques, K-means, Fuzzy C-Means, and hierarchical algorithms are introduced. In addition, the effects of their parameters on the clustering results are investigated based on the three well-known indices including Silhouette index (SIL), Calinski-Harabasz index (CHI), and Davies-Bouldin indicator (DBI). Then a comparison of the mentioned clustering algorithms is presented. Finally, one of the clustering results are correlated to calendar periods.

The rest of the paper is organized as follows: Section II proposed the conducted algorithms for price clustering. An application of clustering algorithms on PUN is presented in Section III. Finally, section IV concludes the paper.

II. METHODOLOGY

In this section, preprocessing of the data is presented in A. Utilized clustering methods are introduced in sub-section B. Three common used CVIs are described in sub-section C.

A. Pre-processing

Gathering and preprocessing of the electricity price data is the first step of price segmentation. Adjustments for daylight saving time changes and data normalization are the simple data pre-processing.

Since the goal is to cluster the daily electricity price curves (365 daily price curves for one year) based on their shapes, a normalization process is performed. In this regard, each hourly price is normalized based on the average price of the given day.

$$P_t^* = \frac{P_t}{\frac{1}{24}\sum_{t=1}^{t=24} P_t} \qquad t \in D \tag{1}$$

where P_t^* is the scaled value of P_t in hour t, and D is the index of each day.

B. Clustering methods

In this subsection, the detailed information of three used well-known clustering algorithms is described.

- K-Means

K-means is a widely used clustering algorithm and it is described as a partitioning method. In this method, the boundary between clusters is fully defined. It is a so-called Crisp clustering. The goal of the algorithm is to minimize the square error criterion, defined as [21]:

$$E = \sum_{k=1}^{K} \sum_{x \epsilon c_k} \|x - c_k\|^2 \tag{2}$$

where K is the number of clusters, x is each observation, and c_k is the mean or center of k^{th} cluster.

- Fuzzy C-means (FCM)

Fuzzy C-means (FCM) is rather similar to K-means clustering, but each point in the dataset has a membership degree to every cluster and the boundary between each cluster is not fully defined. The main goal of the FCM is minimizing the following objective function [18], [21].

$$J_m = \sum_{l=1}^{N} \sum_{k=1}^{K} \mu_{lk}^m \|x_l - c_k\|^2 \tag{3}$$

where N is the number of observations, μ_{lk} is the degree of membership of l^{th} member in k^{th} cluster.

In each cluster, the degree of membership is defined as closeness to the center, the closer to the center, the higher the membership degree, where:

$$\sum_{k=1}^{K} \mu_{lk} = 1 \tag{4}$$

The membership degrees are updated in each step as:

$$\mu_{lk} = \left[\sum_{j=1}^{K} \left[\frac{\|x_l - c_k\|}{\|x_l - c_j\|}\right]^{\frac{2}{m-1}}\right]^{-1} \tag{5}$$

The degree of fuzziness is controlled by m. This value must be greater than 1, with smaller values creating more crisp cluster boundaries.

- Hierarchical clustering

Hierarchical clustering algorithm, which is in the form of trees or dendrograms, gained more attention in the power system field. This method is divided into two categories: an agglomerative (bottom-up) and divisive (top-down)

For N number of observations, the agglomerative method starts with N clusters, and a single observation indicates one cluster. The next step finds the most similar clusters c_i and c_j, then iteratively merged them until the whole tree is formed [22].

To merge each similar pair of clusters, each pair's distance is calculated using a linkage criterion. Some of the most important linkage criteria are Single, Complete, Average, and Ward. More detailed information about the linkage criteria is presented in [4], [23].

C. Clustering Validity Indexes (CVIs)

To evaluate the clustering results, various CVIs including Silhouette index (SIL), Calinski-Harabasz index (CHI), and Davies-Bouldin indicator (DBI) are used. The indexes define as follow [17]:

- *Silhouette index (SIL)*

The mean of the silhouette width of the point x for a given cluster k is defined as:

$$\xi_k = \frac{1}{n_k} \sum_{x \in I_k} \frac{b(x) - a(x)}{\max(a(x), b(x))} \tag{6}$$

where, x denoted of each observation, n_k is the number of data points belonging to cluster k, $a(x)$ is the within-cluster mean distance, $b(x)$ is the mean distance of point M_i to the points of each of the other clusters. I_k is the set of the indices of the observations belonging to the cluster C_k

The global cluster index is defined as:

$$SIL = \frac{1}{K} \sum_{k=1}^{K} \xi_k \tag{7}$$

where, K is the number of clusters

The maximum value of SIL indicates the best result.

- *Calinski-Harabasz index (CHI)*

The Calinski-Harabasz index is defined like this

$$CHI = \frac{BGSS/(K-1)}{WGSS/(N-K)} = \frac{(N-K)}{(K-1)} \frac{BGSS}{WGSS} \tag{8}$$

where, $WGSS$ is the within-cluster dispersion, can also define as sum of the squared distances between the observations of each cluster to the center of the cluster.

Let us denote by$BGSS$ the between-group dispersion. The maximum value of CHI points to the best result.

- *Davies-Bouldin indicator (DBI)*

$$DBI = \frac{1}{K}\sum_{k=1}^{K} M_k = \frac{1}{K}\sum_{k=1,k'\neq k}^{K} \max\left(\frac{\delta_k + \delta_{k'}}{\Delta_{kk'}}\right) \qquad (9)$$

where, δ_k and $\delta_{k'}$ are the mean distance of the points belonging to cluster C_k and to the center of the same cluster and other clusters, respectively.

Let us denote by $\Delta_{kk'}$ the distance between the center of cluster C_k and $C_{k'}$. In DBI, the minimum value indicates the best result.

III. Application of Clustering Algorithms on Italian National Single Electricity Price

In this section, a brief introduction to the data set is presented (Section A).

The process of finding the best parameter of each mentioned algorithm is discussed in section B. Then, three major clustering algorithms including, K-mean, FCM, and hierarchical, are compared, and the outcome of the best clustering result of each algorithm is interpreted Section C

A. *Dataset introduction*

The proposed clustering methods are applied to the daily profiles of Italian National Single Price [1](PUN) for the year 2019. All the data are taken from *Gestore dei Mercati Energetici* (GME) [24] on an hourly basis.

The raw data of daily PUN profile in year 2019 is depicted in Figure 1. There is some variation in PUN in different hour of the day.

B. *Finding best parameter of clustering method*

In this section the process of finding the best parameter of FCM and hierarchical algorithm is described.

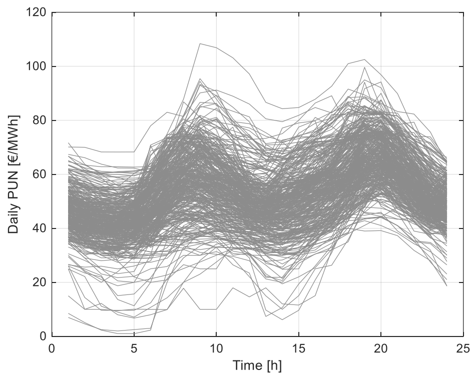


FIGURE 1: DAILY PUN PROFILE IN 2019

- *FCM*

The main parameter of the FCM algorithm is the fuzziness degree (m) which should be tuned by trial-and-error. In data science literature, different values and ranges for fuzzy degree has been suggested [25], [26]. Based on the studies, the range of variation of parameter m is defined between 1.05 to 3.

Due to the random initialization of clustering centers, the outcomes of the clustering methods are slightly different in each execution. So, the clustering is performed, for each value of parameter m, ten times.

[1] Prezzo Unico Nazionale

The best value of fuzziness degree in each number of clusters are listed in Table 1, where the reported values are the average of each CVIs over ten executions of the clustering algorithm.

TABLE 1: THE BEST VALUE OF FUZZINESS DEGREE IN EACH CLUSTER NUMBER

	Best *m* for each CVI		
No.Cluster	**SIL**	**CHI**	**DBI**
2	1.1	1.1	1.1
3	1.1	1.05	1.05
4	1.1	1.1	1.1
5	1.25	1.25	1.3
6	1.1	1.1	1.05
7	1.15	1.2	1.15
8	1.1	1.15	1.1

We use the best value of fuzziness in final comparison.

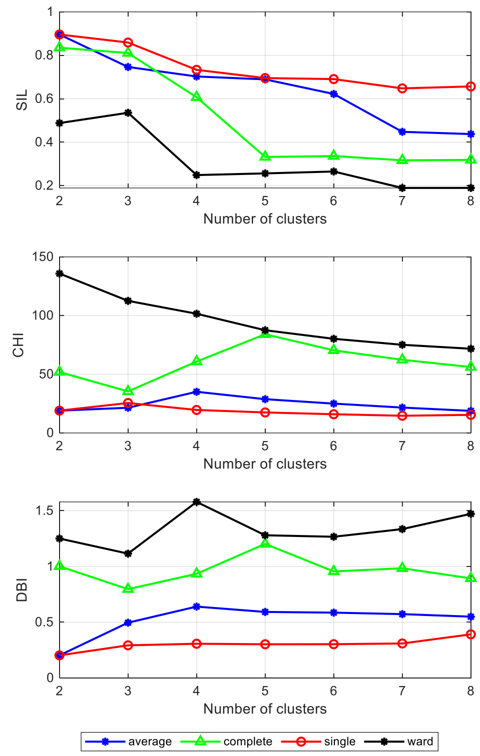


FIGURE 2: COMPARING DIFFERENT LINKAGES OF HIERARCHICAL CLUSTERING ALGORITHM (MAXIMUM VALUE OF SIL AND CHI, AND MINIMUM VALUE OF DBI IS THE BEST RESULT)

- *Hierarchical methods*

Hierarchical algorithm uses linkage criterion for computing the distance between clusters. To find the best linkage criterion, a comparative analysis of four major linkage criteria including Average, Complete, Single, and Ward considering a different number of clusters has been carried out. From Figure 2, adopting Single linkage would result in the best performance in SIL and DBI, but worst performance in CHI indicator. However, Ward linkage shows the best performance on CHI indicator, it shows relatively bad performance on SIL and DBI. Make the final decision based on the results from Figure 2 is difficult.

To have a better insight, the dendrograms of each linkage criteria is depicted in Figure 3. As can be seen, in Single linkage, almost of the observation are grouped in one cluster, but Ward linkage separated all observations in different clusters.

C. Final discussion

In the previous sections, tuning the parameters of FCM and finding the best linkage criterion in the hierarchical algorithm are investigated in detail.

In this section, for the sake of comparison, the clustering methods are performed by varying the number of clusters from 2 to 8. The executions are performed by using the selected parameters from the previous sections. The results are illustrated in Figure 4. Based on the three CVIs, the K-means and FCM algorithm display superior results on two clusters. It is worth mentioning the achieved results for FCM is sensitive to the degree of fuzziness. In this regard, estimating the best degree of fuzziness for FCM is vital. In the hierarchical algorithm with Ward linkage, although the results corresponding to SIL and DBI suggest three clusters is proper, the proposed number of clusters based on the CHI is two clusters.

The final implementation of the three clustering methods on the daily profiles of PUN for the year 2019 is depicted in Figure 5. The number of clusters is set on two. Based on this figure, different clear price patterns in the daily price curves are observable. Basically, the two clusters are the working and holiday day, respectively.

It is worth mentioning that the center of curves classified in cluster 1(working days) started from lower price than cluster 2 (weekend or holidays). In addition, the endpoint of this curve in cluster 1 is lower than cluster 2. Due to the closure of industries on weekends, the midday peak associated with cluster 1 is eliminated in cluster 2.

On the three mentioned algorithms, all the working days and weekends (or holidays) are grouped on clusters 1 and 2, respectively. However, there are different days assigned to the opposite cluster.

By correlating the outcomes of the clustering methods to the calendar periods, Table 2 listed the count of the number of days of the week in each cluster. Between 261 weekdays in 2019, 30 days are grouped in the opposite clusters by using the K-Means algorithm. These numbers for FCM and hierarchical algorithm are 33 and 18, respectively. Although the hierarchical algorithm has better performance in discovering weekdays, this algorithm is incapable of categorizing weekends.

A diligent search in the Italian calendar in 2019 shows that some of the weekdays that grouped in the opposite clusters are holidays. For instance, by analysing the outcomes of the K-Means algorithm and matching the misclassified weekdays by the calendar, we can see 6 days (1 Jan, 22 Apr, 25Apr, 1 May, 15 Aug, and 25 Dec) are public holidays. Other interesting results is related to the period between August 12th to 16th, and August 19th to 23rd. Based on the calendar, in 2019, the August 15th was on Thursday, and a few days prior to and after the public holidays are considered as a summer holiday in Italy. As a result, these periods are classified as a holiday correctly.

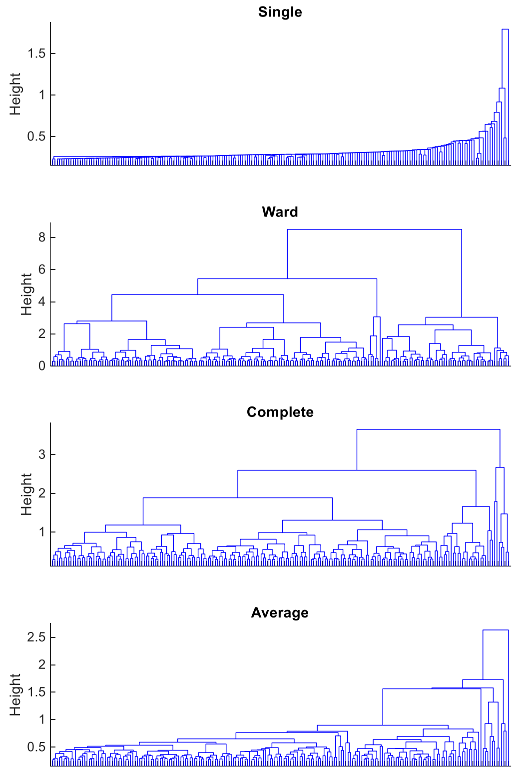


FIGURE 3: DENDROGRAMS OF DIFFERENT LINKAGES OF HIERARCHICAL CLUSTERING ALGORITHM

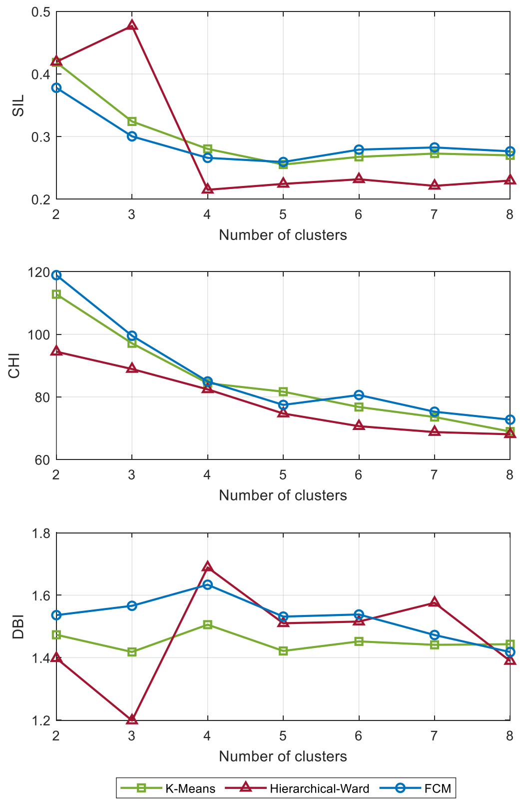
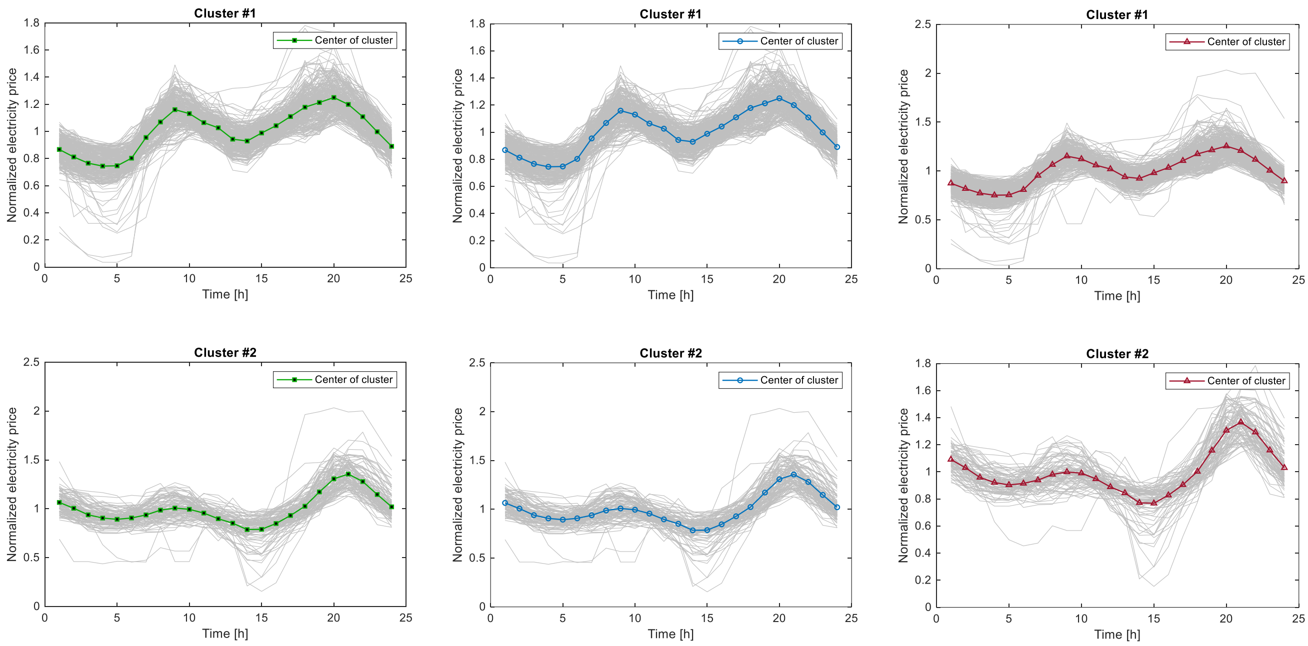


FIGURE 4:COMPARIOSN BETWEEN THREE DIFFERENT ALGORITHM IN DIFFERENT NUMBER OF CLUSTERS



a) K-Means b) Fuzzy C-Means c) Hierarchical

FIGURE 5: FINAL CLUSTER OF EACH ALGORITHM

TABLE 2: COUNT THE NUMBER OF DAYS OF THE WEEK IN EACH CLUSTER

	K-Means		FCM		Hierarchical	
	Cluster1	Cluster2	Cluster1	Cluster2	Cluster1	Cluster2
Sunday	**14**	**38**	**15**	**37**	**20**	**32**
Monday	48	4	48	4	48	4
Tuesday	45	8	44	9	49	4
Wednesday	48	4	48	4	49	3
Thursday	44	8	42	10	49	3
Friday	46	6	46	6	48	4
Saturday	**15**	**37**	**15**	**37**	**18**	**34**

In the obtained clustering results via the K-Means method, 4 working days in March are classified as a weekend or holidays. In addition, April 19th, Good Friday, is segmented as a weekend or holidays group, can be a holiday but it is not a public holiday.

After subtracting the holidays (6 days of public holiday, 9 days of summer holidays in August, and a day holiday in April) from 30 days that are grouped in the opposite cluster, there are 14 days for which no convincing reason can be found for grouping them as a cluster 2

IV. CONCLUSION

In this study, the performance of three main clustering algorithms including K-Means, Fuzzy C-Means (FCM), and Hierarchical algorithm on the daily curves of Italian national single price (PUN) has been investigated.

The main goal was to detect various clear patterns of electricity price which, later, can be utilized for improving the performance of other applications e.g., price prediction, and demand response.

Based on the obtained results the K-means clustering algorithm shown the best performance for identifying clusters of similar price profiles in Italian electricity market.

The investigation on the outcomes of three clustering algorithms, based on the three CVIs including SIL, CHI, and DBI, identified clustering algorithms with two clusters, working days, and weekend or holidays, that yielded the best results in the specific case study.

REFERENCES

[1] D. Keles, J. Scelle, F. Paraschiv, and W. Fichtner, "Extended forecast methods for day-ahead electricity spot prices applying artificial neural networks," *Appl. Energy*, vol. 162, pp. 218–230, 2016.

[2] M. H. Imani, E. Bompard, P. Colella, and T. Huang, "Forecasting electricity price in different time horizons: an application to the Italian electricity market," *IEEE Trans. Ind. Appl.*, 2021.

[3] M. Hosseini Imani, E. Bompard, P. Colella, and T. Huang, “Impact of Wind and Solar Generation on the Italian Zonal Electricity Price,” *Energies* , vol. 14, no. 18. 2021, doi: 10.3390/en14185858.

[4] P. Cichosz, *Data mining algorithms: explained using R*. John Wiley & Sons, 2014.

[5] S. Lin, F. Li, E. Tian, Y. Fu, and D. Li, “Clustering load profiles for demand response applications,” *IEEE Trans. Smart Grid*, vol. 10, no. 2, pp. 1599–1607, 2017.

[6] C. Flath, D. Nicolay, T. Conte, C. Dinther, and L. Filipova-Neumann, “Cluster Analysis of Smart Metering Data. Business & Information Systems Engineering, 4 (1): 31-39.” 2012.

[7] F. Martínez-Álvarez, A. Troncoso, J. C. Riquelme, and J. M. Riquelme, “Partitioning-clustering techniques applied to the electricity price time series,” in *International Conference on Intelligent Data Engineering and Automated Learning*, 2007, pp. 990–999.

[8] K. Zhou, C. Yang, and J. Shen, “Discovering residential electricity consumption patterns through smart-meter data mining: A case study from China,” *Util. Policy*, vol. 44, pp. 73–84, 2017.

[9] Y.-I. Kim, J.-M. Ko, and S.-H. Choi, “Methods for generating TLPs (typical load profiles) for smart grid-based energy programs,” in *2011 IEEE Symposium on Computational Intelligence Applications In Smart Grid (CIASG)*, 2011, pp. 1–6.

[10] D. Roberts and S. F. Brown, “Identifying calendar-correlated day-ahead price profile clusters for enhanced energy storage scheduling,” *Energy Reports*, vol. 6, pp. 35–42, 2020.

[11] R. Li, C. Gu, F. Li, G. Shaddick, and M. Dale, “Development of low voltage network templates—Part I: Substation clustering and classification,” *IEEE Trans. Power Syst.*, vol. 30, no. 6, pp. 3036–3044, 2014.

[12] O. Motlagh, G. Foliente, and G. Grozev, “Knowledge-mining the Australian smart grid smart city data: A statistical-neural approach to demand-response analysis,” in *Planning support systems and smart cities*, Springer, 2015, pp. 189–207.

[13] I. B. Sanchez, I. D. Espinós, L. M. Sarrión, A. Q. López, and I. N. Burgos, “Clients segmentation according to their domestic energy consumption by the use of self-organizing maps,” in *2009 6th International Conference on the European Energy Market*, 2009, pp. 1–6.

[14] L. A. P. Júnior *et al.*, “Unsupervised non-technical losses identification through optimum-path forest,” *Electr. Power Syst. Res.*, vol. 140, pp. 413–423, 2016.

[15] K. Li, Z. Ma, D. Robinson, and J. Ma, “Identification of typical building daily electricity usage profiles using Gaussian mixture model-based clustering and hierarchical clustering,” *Appl. Energy*, vol. 231, pp. 331–342, 2018.

[16] H. Teichgraeber and A. R. Brandt, “Clustering methods to find representative periods for the optimization of energy systems: An initial framework and comparison,” *Appl. Energy*, vol. 239, pp. 1283–1293, 2019.

[17] B. Desgraupes, “Clustering indices: University of paris ouest-lab modal’x; 2013,” *URL f tp//apache. cs. uu. nl/mirror/CRAN/web/packages/clusterCrit/vig nettes/clusterCrit. pdf*, pp. 1–34.

[18] G. Chicco, R. Napoli, and F. Piglione, “Comparisons among clustering techniques for electricity customer classification,” *IEEE Trans. power Syst.*, vol. 21, no. 2, pp. 933–940, 2006.

[19] V. Figueiredo, F. Rodrigues, Z. Vale, and J. B. Gouveia, “An electric energy consumer characterization framework based on data mining techniques,” *IEEE Trans. power Syst.*, vol. 20, no. 2, pp. 596–602, 2005.

[20] T.-S. Xu, H.-D. Chiang, G.-Y. Liu, and C.-W. Tan, “Hierarchical K-means method for clustering large-scale advanced metering infrastructure data,” *IEEE Trans. Power Deliv.*, vol. 32, no. 2, pp. 609–616, 2015.

[21] T. Zhang, G. Zhang, J. Lu, X. Feng, and W. Yang, “A new index and classification approach for load pattern analysis of large electricity customers,” *IEEE Trans. Power Syst.*, vol. 27, no. 1, pp. 153–160, 2011.

[22] N. Rajalingam and K. Ranjini, “Hierarchical clustering algorithm-a comparative study,” *Int. J. Comput. Appl.*, vol. 19, no. 3, pp. 42–46, 2011.

[23] F. Murtagh and P. Legendre, “Ward’s hierarchical agglomerative clustering method: which algorithms implement Ward’s criterion?,” *J. Classif.*, vol. 31, no. 3, pp. 274–295, 2014.

[24] “GME - Gestore dei Mercati Energetici SpA.” https://www.mercatoelettrico.org/En/Default.aspx (accessed Dec. 10, 2020).

[25] A. Rajabi, M. Eskandari, M. J. Ghadi, L. Li, J. Zhang, and P. Siano, “A comparative study of clustering techniques for electrical load pattern segmentation,” *Renew. Sustain. Energy Rev.*, vol. 120, p. 109628, 2020.

[26] K.-L. Wu, “Analysis of parameter selections for fuzzy c-means,” *Pattern Recognit.*, vol. 45, no. 1, pp. 407–415, 2012.