AIMS Energy

AIMS Energy, 10(6): 1146–1164.
DOI: 10.3934/energy.2022054
Received: 08 August 2022
Revised: 28 October 2022
Accepted: 03 November 2022
Published: 11 November 2022

http://www.aimspress.com/journal/energy

Research article

On the detection of patterns in electricity prices across European countries: An unsupervised machine learning approach

Dimitrios Saligkaras and Vasileios E. Papageorgiou*

Department of Mathematics, Aristotle University of Thessaloniki, Thessaloniki 54124, Greece

* **Correspondence:** Email: vpapageor@math.auth.gr.

Abstract: The year 2022 is characterized by a generalized energy crisis, which leads to steadily increasing electricity prices around the world, while the corresponding salaries remain stable. Therefore, examining trends in electricity prices relative to existing income levels can provide valuable insights into the overpricing/underpricing of energy consumption. In this article, we examine the tendencies of 35 European countries according to their national kWh prices and the average household incomes. We use a series of established clustering methods that leverage available information to reveal price and income patterns across Europe. We obtain important information on the balance between family earnings and electricity prices in each European country and are able to identify countries and regions that offer the most and least favorable economic conditions based on these two characteristics studied. Our analysis reveals the existence of four price and income patterns that reflect geographical differences across Europe. Countries such as Iceland, Norway, and Luxembourg exhibit the most favorable balance between prices and earnings. Conversely, electricity prices appear to be overpriced in many southern and eastern countries, with Portugal being the most prominent example of this phenomenon. In general, average household incomes become more satisfactory for European citizens as we move from east to west and south to north. In contrast, the respective national electricity prices do not follow this geographical pattern, leading to notable imbalances. After identifying significant cases of inflated prices, we investigate the respective causes of the observed situation with the aim of explaining this extreme behavior with exogenous factors. Finally, it becomes clear that the recent increase in energy prices should not be considered as a completely unexpected event, but rather as a phenomenon that has occurred and developed gradually over the years.

Keywords: clustering algorithms; electricity prices; Partition Around Medoids; hierarchical clustering; household incomes; unsupervised machine learning

1. Introduction

The first half of 2022 has been associated in many people's minds with a time of increased expensiveness and uncontrollably rising fuel and electricity prices. There is no doubt that energy prices affect all households. A household's ability to pay its energy-related bills is determined by two main factors, energy prices and the salary of the average consumer. Therefore, overpriced energy bills that are disproportionate to consumers' wages significantly worsen the overall standard of living. But how can it be said that a country's citizens are burdened by electricity prices, while in other countries such burden does not exist?

Unsupervised methodologies such as clustering algorithms, allow countries to be categorized in terms of household incomes and national kWh price levels. Unlike supervised learning, where each object (data point, image, text, etc.) should be given an appropriate label [1,2], unsupervised methods can operate solely based on information provided by the independent variables exclusively. Such categorization makes any comparison feasible. In recent decades, the problem of cluster analysis has been thoroughly studied and various algorithms have been proposed. Although, the way clustering is approached varies greatly from algorithm to algorithm [3,4]. Fraley and Raftery [5] proposed a division of the different clustering techniques into hierarchical and partitioning. However, there are other techniques based on the density of observations or the application of statistical models, leading to the addition of more clustering categories to the already existing ones [6].

Matuszewska-Janica et al. [7] performed an analysis based on K-means between European countries to show the differences between households in terms of sustainable energy development. Gostkowski et al. [8] also used the K-means algorithm and other agglomerative methods to assess energy consumption, but only for the Visegrad Group countries, using data from 1990 to 2018. Poyrazoglu [9] attempted to define electricity price zones, but only for the Turkish market. For the latter, K-means and queen/rook spatial constraint clustering were applied. In addition to the above articles, Mart ńez et al. [10] had a quite similar approach using the conventional and the fuzzy K-means.

Furthermore, Verbič et al. [11] investigated the impact of residential electricity prices on energy intensity in Europe using a linear regression model, while Gil-Alana et al. [12] studied the relationship between energy consumption and prices in Spain and Portugal, utilizing linear time series analysis, namely autocorrelation indicators and autoregressive models. Takentsi et al. [13] examined the existence of causal relations between electricity prices and economic performance in South Africa for the period 1994–2019. Their findings reveal that electricity prices have a significant negative impact on economic growth in the long and short run. Shah et al. [14] presented a comparison of multivariate vector autoregressive models where their parameter estimation is based on least squares, Lasso, Ridge or Elastic-net regression. The forecasting efficiency of these four models is tested on electricity demand and price time series between 2013 and 2017. Also, Wang and Li [15] examined the impact of electricity prices on power-generation structures, showing that an increase of electricity prices can not only improve the efficiency of power plants, but also propel firms to invest more in renewable energy plants. Moreover, the usage of other techniques based on indicators like the Pearson or Spearman correlations, could provide valuable information about the investigated phenomenon [16].

Zhou and Zheng [17] used artificial neural networks to forecast building demand, since efficient demand forecasting and advanced demand-side controls are necessary to improve building energy flexibility. Zhou et al. [18] provided an overview of passive and active phase change materials, integrated building energy systems with advanced machine learning-based climate-adaptive designs and uncertainty-based optimizations. In addition, Liu et al. [19] developed an uncertainty energy design for net-zero energy communities. According to the authors, this design, which involves hydrogen and battery storage vehicles, provides important guidance for advancing net-zero energy and achieving carbon neutrality for integrated building and transportation sectors by 2050. All previous studies feature innovative elements and make valuable contributions to the field of energy studies.

In this article, special emphasis is placed on the study of kWh prices compared to mean household incomes across Europe. Unlike the aforementioned studies, our work aims to include income data from 35 European countries in the discussion, leading to more accurate and reliable conclusions about the existing situation. As a result, we arrive at important observations about the balance between kWh prices and family earnings in each country, distinguishing cases where the given conditions seem to be clearly in favor or to the disadvantage of the average consumer.

The reliability of our results is significantly increased due to the implementation of several established clustering techniques, such as K-means++, PAM (Partition Around Medoids), CLARA (Clustering Large Applications), DBSCAN (Density Based Clustering of Applications with Noise), and UPGMA (Unweighted Pair Group Method using arithmetic Average). Most of the aforementioned studies are limited to the exploration of only one method. The reason for using such a variety of techniques is to derive accurate results, as different clustering algorithms may yield different outcomes depending on how they work. In addition, our approach helps to identify notable cases of overvaluation, leading to an investigation of the respective causes behind the observed situation, aiming to justify this extreme behavior, which may be influenced by exogenous factors.

In our analysis, we come across characteristic geographical patterns based on national electricity prices and household earnings, which allow us to draw valuable conclusions about the existing situation in Europe. Finally, according to the produced clustering results, we find the general tendency that more and more European countries are systematically classified into clusters representing higher kWh price standards during the period 2007–2018. This phenomenon suggests that the current energy crisis has its roots before the recent Russian-Ukrainian war, which is considered to be the main culprit behind the rising energy costs.

The paper is organized as follows. In section 2, we present the main features as well as the steps of the algorithm for each of the considered clustering techniques. In section 3, we investigate the existence of patterns in Europe in terms of national electricity prices and corresponding average incomes using five established clustering methods. Finally, in sections 4 and 5, we summarize and discuss the obtained results, focusing on general upward price trends. Hence, we highlight cases of extreme overpricing and finally draw valuable conclusions about the current situation in Europe with respect to kWh prices relative to average incomes.

2. Materials and methods

2.1. Hierarchical clustering

Agglomerative methods, belonging to the class of hierarchical clustering algorithms, follow a bottom-up approach where the initially created clusters contain only one element [20]. During each step of the process, the initial clusters merge, resulting in groups containing more and more elements [21]. The agglomerative procedure terminates until all objects are placed in a single cluster. Hierarchical methods usually lead to the creation of tree diagrams that help draw useful conclusions about the appropriate number of clusters [22]. The decision of which clusters to join or separate is defined based on a similarity measure, which combines the distance of elements with a linkage criterion that determines the similarity of two groups. The three most common linkage criteria between two clusters A and B are [23]:

1. Complete-linkage, defined as $\max\{d(x,y): x \in A, y \in B\}$ and representing the maximum distance between the elements of A and B.
2. Single-linkage, defined as $\min\{d(x,y): x \in A, y \in B\}$ and representing the minimum distance between the elements of A and B.
3. Average-linkage, defined as $\frac{1}{|A||B|}\sum_{x\in A}\sum_{y\in B} d(x,y)$ and representing the average distance between the elements of A and B.

The process stops when the number of classes becomes sufficiently small [24]. Lance and Williams [25], proposed an equation that with appropriate adjustments represents any hierarchical model. For the purposes of our analysis, we emphasize the UPGMA clustering methodology corresponding to the average-linkage case.

The group average method—also called UPGMA—is defined as the average distance between all examined classes. The distance between a merged class $C_i \cup C_j$ and a third C_k, results from

$$D(C_i \cup C_j, C_k) = \frac{|C_i|}{|C_i| + |C_j|} D(C_i, C_k) + \frac{|C_j|}{|C_i| + |C_j|} D(C_j, C_k) \tag{1}$$

while for classes C and C' we have

$$D(C, C') = \frac{1}{|C||C'|} \sum_{x\in C, y\in C'} d(x,y) \tag{2}$$

2.2. K-means

Given an initial set of k centers $m_1^1, m_2^1, \dots, m_k^1$, the algorithm works based on the following steps [26]:

Step 1. Each observation must be included in the cluster of its nearest medoid. Usually, the appropriate metric that defines proximity is the squared Euclidean distance,

$$S_i^t = \left\{x_p : \left|x_p - m_i^t\right|^2 \le \left|x_p - m_j^t\right|^2, \qquad 1 \le i, j \le k\right\} \tag{3}$$

where each observation x_p is inserted only in cluster S^t. Other commonly used distances are the Chebyshev and Minkowski [27].

Step 2. The medoids of the new clusters are calculated as

$$m_i^{t+1} = \frac{1}{|S_i^t|} \sum x_j, \tag{4}$$

where $|S_i^t|$ denotes the number of elements that have joined the S_i class during the *t*-th iteration. During the algorithm's operation, we aim to minimize the value of the within clusters sum of squares function (WCSS), presented by

$$wcss = \sum_{i=1}^{k} \sum_{x_j \in S_i} \left|x_j - m_i\right|^2. \tag{5}$$

This procedure leads to diversified results based on the preselected number of clusters [28].

2.3. K-means++

K-means++ attempts to enhance the k initial medoids selection procedure offered by the naive K-means. An object is randomly selected from the dataset to be the first medoid. Each item is selected based on the uniform distribution. The remaining objects are selected as medoids with a probability proportional to the distance from the nearest, already chosen medoid [29]. This process continues until k centers are selected. Briefly, the algorithm consists of the following 4 steps:

Step 1. Select a medoid c_1 from dataset X according to the uniform distribution.

Step 2. Choose a medoid c_i from X, with probability $\frac{D(x)^2}{\sum_{x \in X} D(x)^2}$.

Step 3. Repeat the previous step until k centers are selected.

Step 4. Proceed as dictated by the K-means algorithm.

Applications have shown that K-means++ outperforms the conventional K-means, highlighting the importance of efficient medoid initialization. Selecting the initial centers requires k additional scans of the dataset. However, this process does not significantly increases the convergence time of K-means++, as a proper medoid initialization usually leads to faster convergence times.

2.4. PAM

The PAM algorithm identifies objects called centroids, which are important for initializing the iterative procedure. The centroids are placed in a set S called the set of selected objects. If O is the set of all objects, we define the set of unselected elements $U = O - S$. This method minimizes the ambiguity of the objects with the nearest selected element (centroid) by using the Euclidean or, in certain cases, the Manhattan distance. The algorithm consists of two main phases [30].

Phase 1. (BUILD). A collection of k objects is selected to create the initial set S.

Phase 2. (SWAP). In this phase, we improve the initial clustering quality by swapping selected and unselected elements. We define quantities D_p and E_p, which denote the dissimilarity between an object

p and its 1st and 2nd nearest neighbors in the set S, respectively. The execution of the PAM is based on the following steps:

Step 1. The object with the minimum total distance relative to all other objects is placed in the set S, representing the most central element of O.

Step 2. An object $i \in U$, is randomly selected as a candidate member in S.

Step 3. For an object $j \in U - \{i\}$, calculate quantity D_j. In case $d\,(i,j) < D_j$, then j contributes to the selection of object i.

Step 4. Consider $C_{ij} = max(D_j - d\,(i,j), 0)$ and calculate the quantity $g_i = \sum_{j \in U} C_{ij}$. Finally, select the object i that maximizes the quantity g_i.

Step 5. Sets S and U are updated after each iteration.

These steps are performed until k elements are selected for the set S. During the swapping phase, we aim to improve the clustering quality. The necessary steps are:

Step 1. Calculate K_{ijh} considering the following two cases

- if $d(j,i) > D_j$ then $K_{ijh} = \min\left(d_{jh} - D_j, 0\right), \quad (i,h) \in S \times U,\ j \in U - \{h\}$,

- if $d(j,i) = D_j$ then $K_{ijh} = \min\left(d_{jh}, E_j\right) - D_j,\ (i,h) \in S \times U,\ j \in U - \{h\}$.

Step 2. Calculate $T_{ih} = \sum K_{jih}\,,\ j \in U$.

Step 3. Select the pair $(i,h) \in S \times U$ that minimizes T_{ih}.

Step 4. If $T_{ih} < 0$, perform the substitution of the two elements and return to step 1. The algorithm terminates when all T_{ih} become positive.

2.5. CLARA

CLARA, in contrast with PAM, was designed to manage large datasets [30]. Instead of finding representative objects for the entire dataset, it takes a sample of it and applies PAM to find its centroids. To enhance the algorithm's accuracy, the sampling process is performed several times. Finally, we choose the clustering that minimizes the mean difference of dataset's elements. The steps for CLARA algorithm are as follows:

Step 1. Select a sample of $40 + 2k$ objects using simple random sampling and then perform the PAM algorithm to find the k centers of the sample.

Step 2. For each item of the set U, determine the most similar centroid.

Step 3. Calculate the mean difference of the dataset. If it is smaller than the already existed ones, set this difference as the new minimum. Then, consider the centers that have been found in step 2 as the representative set.

Step 4. Return to step 1 for the next iteration.

The algorithm returns the clusters resulting from the sample with the lowest mean heterogeneity.

2.6. DBSCAN

DBSCAN was proposed by Ester et al. in 1996 [31] and its operation is based on the density of the observations. Specifically, this methodology places together observations that have many neighboring elements, while pointing out items (as irregularities) that their nearest neighbors are far away. To define the neighborhood around a reference point it is necessary to determine a positive radius ε, rendering any other point located inside the supersphere as an adjacent one. Also, a quantity $minPts$ should be specified, indicating the minimum number of points for which an area is considered dense.

A point p is called a *core point* if in a radius ε around it, there are at least $minPts$ points including p. A point q is called directly reachable from a core point p if the distance from q to p is less than ε. A point q is called a *reachable point* from a point p if there exists a path $p_1, \ldots . p_n$, where $p_1 = p$ and $p_n = q$, while every p_{i+1} is directly reachable from p. Moreover, p is *tightly connected* to q, if there is an element o that is a close approximation of p and q. Outliers or noise are the elements that do not fall into any of the abovementioned categories [31].

If p is a core point, it forms a cluster together with the points representing its approximations. Namely, the algorithm selects arbitrarily a starting point p and scans its neighborhood. When the neighborhood is sufficiently dense, the formation of a cluster C initiates, including all points belonging in the vicinity of p. Then, we explore the neighborhood of each participating point q. When the neighborhood of these points contains more than $minPts$ elements, then all neighbors of q that are not already in cluster C, are added. After the completion of this iterative process, the algorithm characterizes the remaining points as noise.

Let d be the distance of a point p from its k-th closest neighbor. Then, the d neighborhood $Nd(\epsilon)$ of p contains exactly $k + 1$ elements. For a given value of k, we define the $k - dist$ function that maps each observation to its distance from the k-th nearest neighbor. These points are placed in descending $k - dist$ order, resulting in the creation of a declining curve.

Consider the value ε to be equal to $k - dist\ (p)$ for some arbitrary element p and $minPts = k$. Our goal is to find the part of the curve that displays a plateau. Points corresponding to the right part of the plateau are considered outliers, while those on the left should be added to the clusters. In practice, various applications have shown that the $k - dist$ graphs for $k > 4$ become similar [31]. Therefore, in the results section we explore the corresponding graphs for $k = 3, 4, 5$. More information on the operation of DBSCAN or any of the abovementioned methodologies can be found in [32–34].

3. Results

In this section, we examine annual Eurostat’s recordings of electricity prices and household incomes based on the abovementioned clustering methodologies. The dataset studied, contains electricity prices (€/kWh) taking into account the additional taxes and levies imposed by the governments of the countries under consideration. For each country there are two values for the kWh price, one for each semester. The average of the two semesters is considered the kWh price per year. The 27 of the 35 participants included in our analysis are members of the European Union (EU), with Turkey, Norway, Iceland, the United Kingdom, Serbia, Northern Macedonia, Montenegro, and Kosovo representing the remaining countries in our study.

Another important aspect to consider, is the corresponding average household income in these countries. Particular attention is paid to the simultaneous investigation of these two characteristics, with the aim of revealing patterns around Europe. We provide the links for the income data and electricity prices under all taxes and levies. The interested reader can find the utilized household incomes for 2018 in the link https://ec.europa.eu/eurostat/databrowser/view/ilc_di16/default/table?lang=en and the corresponding electricity prices during the time period 2007–2018 in the link https://ec.europa.eu/eurostat/databrowser/view/NRG_PC_204__custom_3683549/default/table?lang=en.

The original purpose of this article is to classify European countries and to draw conclusions about electricity prices as a function of the income of an average family. Using the clustering algorithms discussed earlier, we aim to find out in which countries the distribution of average incomes does not match the respective kWh prices, while looking for specific factors that cause such imbalances. At the same time, we explore the evolution of electricity prices for the period 2007–2018, showing a general tendency in energy prices observed in most of the participating European countries.

Our analysis is based on the results of K-means++, PAM, CLARA, DBSCAN and UPGMA. We avoid using algorithms such as OPTICS due to the small number of observations. This limitation is associated with the lack of dense regions, which may lead to unreliable results [35]. We also avoid using the conventional K-means algorithm due to its limited efficiency. Finally, we do not examine the results of the Gaussian mixed models, as they are best suited for a larger number of observations, since the accuracy of the estimation of the ellipsoids requires a satisfactory element cardinality [36]. According to the plots of within clusters sum of squares for kWh prices and household earnings (Figure 1), we do not observe significant changes in the WCSS for a number of clusters greater than 4. Hence, the choice of 4 clusters will be followed throughout this section.

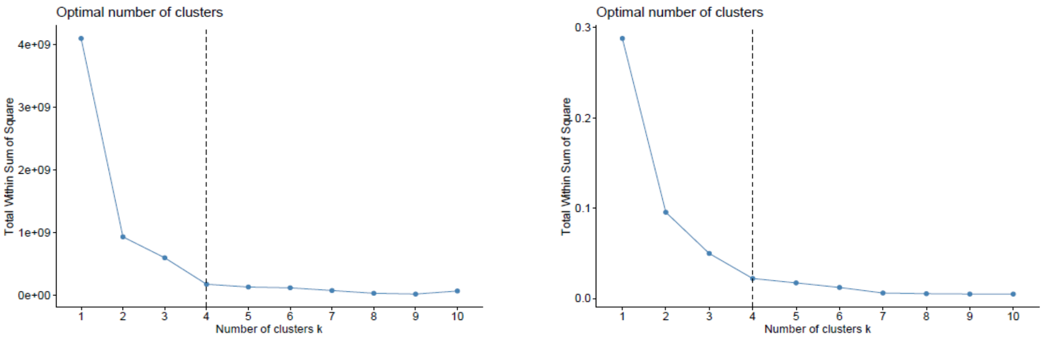

Figure 1. WCSS diagram for the electricity prices and mean household incomes.

First, we consider the grouping of the 35 countries studied, based on the income of an average family using the K-means++ algorithm (Figure 2). The red cluster includes the countries with the highest incomes, the blue cluster includes the countries with the second highest incomes, while the countries with the lowest average household earnings are shown in green. There is an obvious distinction between the western and eastern parts of Europe, where the eastern countries are classified in the yellow and green clusters, representing the cases whose citizens receive the least prosperous wages.

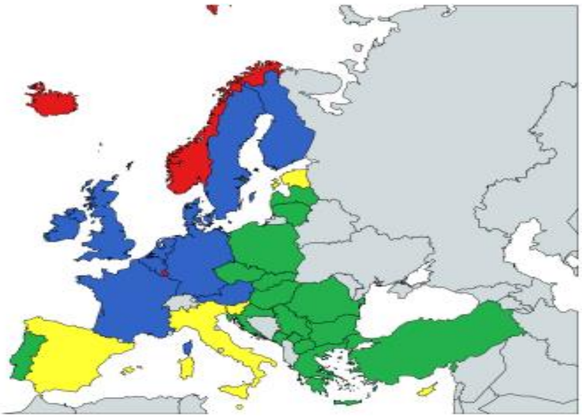

Figure 2. Clustering results of average household incomes of European countries by K-means++.

The green cluster contains incomes between 2059–9346 euros, the yellow cluster has incomes between 10524–16844 euros, while the blue cluster is characterized by incomes between 21464–30104 euros and the red cluster by incomes between 34472–39918 euros. We note that the group of countries with the most satisfactory average family earnings is quite small and includes only Norway, Iceland, and Luxembourg. An interesting observation is the inclusion of Estonia in the yellow group, corresponding to higher average incomes compared to most eastern European countries. There seems to be a clear distinction between the northern and Mediterranean countries, with the latter classified in the green and yellow groups, showing a decreasing trend in salaries as we move from north to south. Namely, participants of the green cluster are Serbia, Greece, Turkey, Poland, Czechia, etc., while the yellow group contain countries such as Italy, Spain and Estonia. Another important observation is the separation of Norway from the other 2 Scandinavian countries and Finland, which are now included in the blue cluster, suggesting that Norway offers quite more generous salaries than its neighbors.

The application of the UPGMA procedure, provide equivalent results with the K-means++ algorithm, where the only exception is the case of Denmark that is included in the red cluster, corresponding to the group of highest incomes. Now Denmark resembles the behavior of Norway, differentiating it from Sweden and Finland.

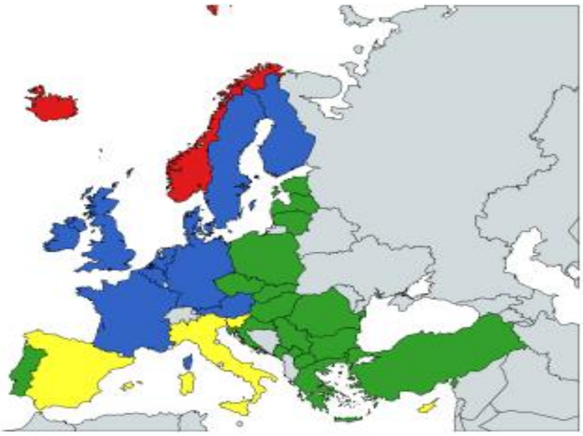

Figure 3. Clustering of European countries based on the average family incomes using PAM or CLARA.

The results of the application of CLARA and PAM are shown in Figure 3. The color interpretation remains the same, while this color selection is also used for the clustering of the national electricity prices, i.e., the highest national kWh prices are assigned red color, etc. The green cluster contains family earnings between 2059–10524 euros, the yellow cluster includes earnings between 13244–16844 euros, while the blue cluster is characterized by earnings between 21464–30104 euros and the red group by earnings between 34472–39918 euros. The results of these two algorithms are similar, a fact that is quite expected based on the operation of these methods. Moreover, there are minor differences compared to the clustering proposed by K-means++, where Estonia is now also included in the green cluster, along with the rest of the Baltic and eastern countries. The results of Figure 3, show a clear distinction between southeastern and northwestern countries in terms of the earnings of an average family.

At this point, we proceed to apply the same algorithms to national electricity prices. The results after applying K-means++ are shown in Figure 4. The color interpretation is similar to the interpretation of the graph of mean incomes. There is no significant pattern in electricity prices depending on the location of the countries studied, as there are cases with relatively high kWh prices in both western and eastern Europe. The red group contains countries such as Germany, Denmark and Belgium, which represent the cases with the most expensive electricity in Europe during 2018. On the other hand, the eastern and southern participants of the analysis are included in the yellow and green groups.

Countries like Greece, Poland, Romania, Slovakia, Hungary, Czech Republic, Croatia and Latvia belong to the group with the third most expensive kWh (yellow cluster), while the same algorithm placed them in the group with the least prosperous salaries (green cluster). Also, countries such as Italy and Spain are now classified in blue cluster of the second most expensive energy prices, while they were classified in the yellow group according to the mean incomes of their citizens. Portugal is particularly noteworthy, as it is the only country included in the group of less generous salaries and at the same time in the group of the second most expensive kWh.

On the other hand, there are also cases where the electricity prices and the average family earnings seem to be balanced. For instance, countries like Serbia, North Macedonia, Kosovo, Turkey, Latvia and Bulgaria are placed in the green cluster. These countries are characterized by lower incomes compared to participants in western and northern Europe, although the respective energy prices do not impose significant financial burden on the average household. Therefore, these countries provide more favorable conditions to their citizens, compared to cases like Greece, Poland, Romania, Croatia etc., where the kWh value is disproportionate to the consumer’s financial capabilities.

Figure 5 shows the classification derived from PAM and CLARA, as they provided similar results. The only exceptions between these methods and K-means++ are France and Hungary, which are included in clusters with higher electricity prices. Looking at the results of all the algorithms used, cases such as Finland and France seem to offer cheaper kWh prices compared to the income of their citizens. This phenomenon is reinforced in the cases of Iceland, Norway and Luxembourg, which are simultaneously found in the group of the wealthiest salaries and the third group of electricity prices. Finally, the UPGMA method provides similar results to K-means++ with the difference that the UK is now classified in the yellow cluster.

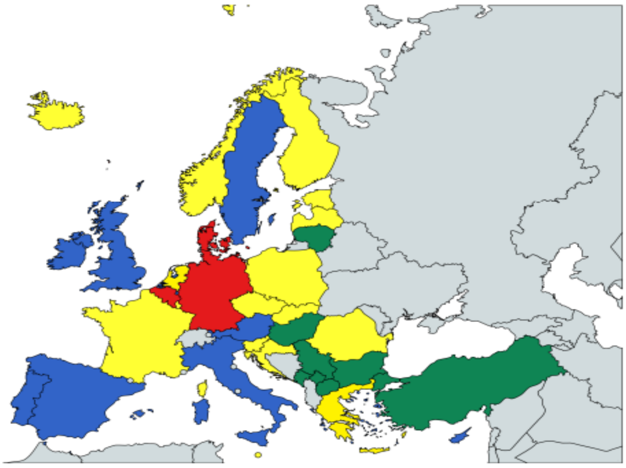

Figure 4. Clustering of national kWh prices in 2018 according to K-means++.

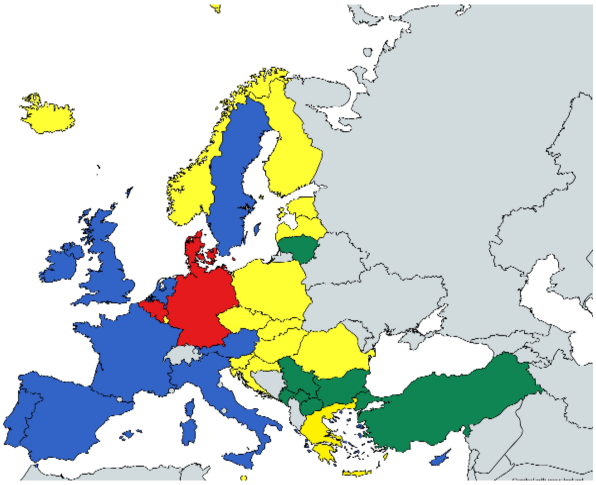

Figure 5. Clustering of European countries based on their electricity prices in 2018 using PAM or CLARA.

In Figure 4, the green cluster includes values between 0.0636–0.1121 €/kWh and the yellow cluster includes prices between 0.1296–0.1712 €/kWh. Moreover, the blue group contains prices between 0.1774–0.2454 €/kWh and the red group between 0.2880–0.3125 €/kWh. On the other hand, the green group of Figure 5, represents prices from 0.0636 to 0.1097 €/kWh, the yellow group from 0.1121 to 0.1681 €/kWh, the blue group from 0.1712 to 0.2454 €/kWh and the red group from 0.2880 to 0.3125 €/kWh. The lowest electricity prices are provided by Kosovo with 0.0636 €/kWh and the highest by Denmark with 0.3125 €/kWh.

An important observation regarding kWh prices for all displayed algorithms is that Denmark and Germany consistently provide the most expensive electricity, at least for 2018. The explanation for this phenomenon is that Denmark ranks first based on total electricity taxes and levies. That is, 70% of the final kWh cost consists of taxes. The reason for this level of taxation is the government's decision since 2012, to support investments that will gradually move the country away from traditional electricity generation methods and towards less polluting alternatives. The main goal of the proposed plan is for half of Denmark's electricity generation to come from renewable energy sources by 2020.

Germany is on the same wavelength, where taxes account for half of the final price of a kilowatt-hour, as the government has decided to significantly reduce the use of lignite for electricity generation since 2000. Therefore, K-means++, PAM, CLARA and UPGMA have identified and separated these 2 cases, as they are the 2 countries with the highest additional burdens on electricity prices. Another country whose citizens seem to pay a lot for electricity is Belgium. In 2018, complications arose at Belgian nuclear power plants, which meant that electricity had to be imported from neighboring France, leading to an abrupt increase in electricity prices. In countries such as Greece, Latvia, Poland, Slovakia, Romania and especially Portugal, the price level seems to be much higher compared to the mean income levels.

Now we turn to the exploration of electricity prices over the period 2007–2018. First, the countries are divided into clusters for each year of the period 2007–2017. Each country is then ranked in the cluster to which it belonged for the most years. This approach helps to examine all 35 European countries, as during the period 2007–2017 we encounter several missing values regarding kWh prices. Missing values are observed in many countries such as Italy, Iceland, Lichtenstein and the Balkan countries Kosovo, Montenegro, North Macedonia, and Serbia. Namely, recordings for Italy start from the second semester of 2010, recordings for Kosovo and Serbia initiate from 2013, recordings in Lichtenstein begin in 2014 etc. Therefore, a grouping of the countries studied using the entire 2007–2017 in one clustering attempt is not possible due to the aforementioned limitations. Presented missing values can only be replaced using an extrapolation methodology, which is highly unreliable due to the short length of the dataset's time series.

Then, a comparison can be made with the clustering for the year 2018. We note that we present only the implementation of K-means++, since the produced results are almost identical to those of PAM, CLARA and UPGMA (Figure 6). Countries like Iceland, the United Kingdom, Estonia, Romania, Hungary and Belgium are classified in higher price clusters in 2018 than in most of the 2007–2017 period. At the same time, the only country with lower electricity prices in 2018 compared to the 2007–2017 period is Norway. Perhaps this is an indication of generally rising kWh prices in Europe during 2018.

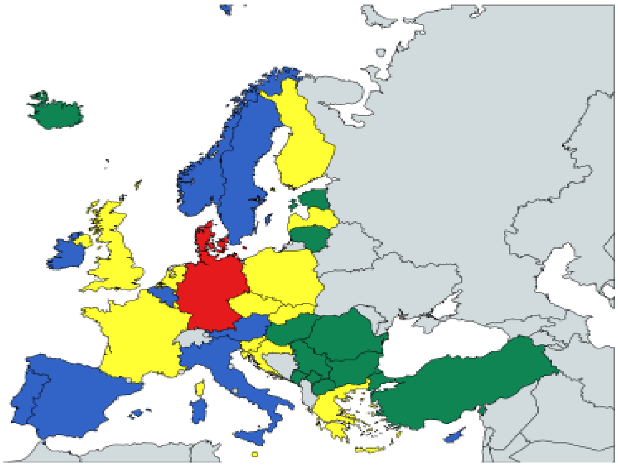

Figure 6. Clustering using K-means++ for the time period of 2007–2017.

The only two countries that remain consistently in the red group of the highest electricity prices are Denmark and Germany. This fact is consistent with the comments we made earlier about the additional taxes on the price of electricity in the early 2010s and 2000s. Moreover, the increase in the kWh price in Belgium in 2018—due to disruptions in the national nuclear power plants—is also confirmed by this part of the analysis. In general, Belgium is classified in the blue cluster for the period 2007–2017. Looking separately at the kWh price of 2018, Belgium belongs to the red cluster of countries with the most expensive kWh. Thus, the nuclear power plant incidents in 2018 raised the prices to a level that is not representative for the country.

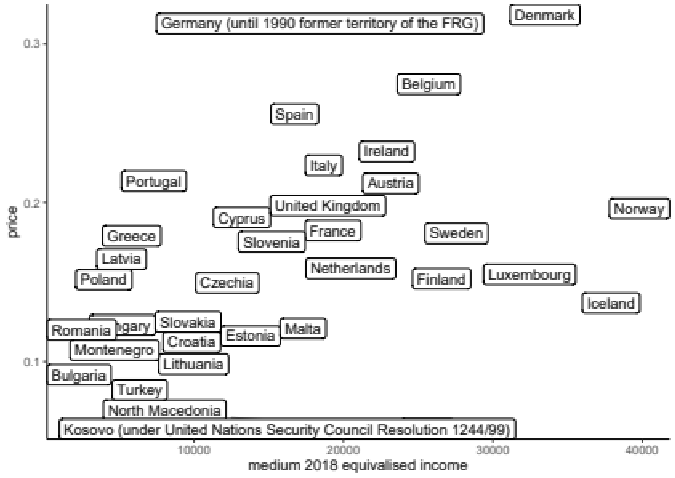

Figure 7. Two-dimensional representation of the household incomes of European countries in relation to the respective kWh value.

Finally, we group the 35 examined countries regarding average family earnings and corresponding national kWh prices simultaneously. In Figure 7, we display a two-dimensional graph presenting the examined European countries based on their electricity prices per kWh in parallel with household incomes. This visual representation validates the aforementioned comments about the imbalances between prices and earnings. Countries that are close to the diagonal provide relatively balanced conditions to their citizens. On the other hand, as we move away for the diagonal line countries either provide favorable salaries or overpriced electricity. For example, countries such as Norway, Luxembourg, Iceland, Sweden and Finland show higher incomes compared to the electricity costs, while others like Portugal, Greece, Latvia etc. display disadvantageous conditions for their inhabitants.

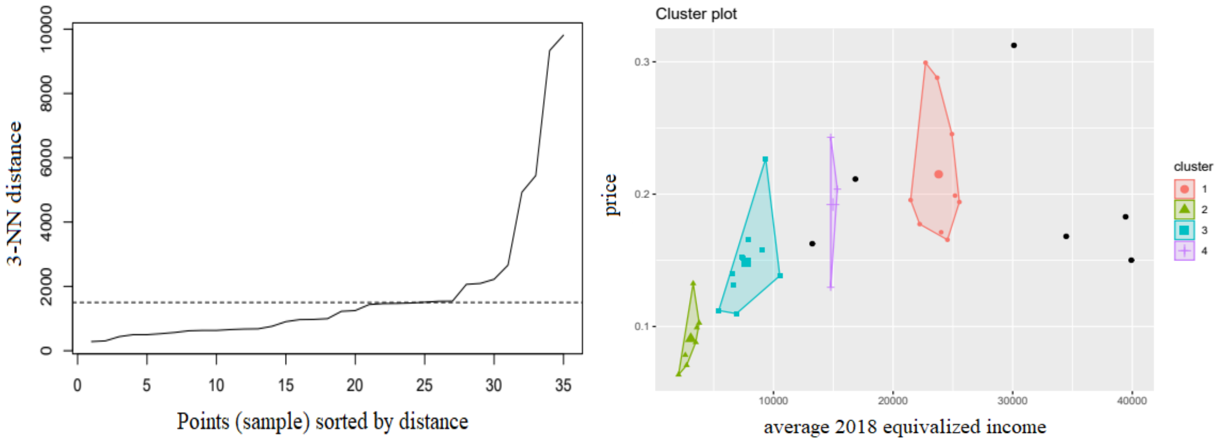

Figure 8. 3-NN distance graph and clustering of European countries regarding the average household earnings and the respective electricity prices using DBSCAN.

In this part we utilize DBSCAN aiming to identify irregular cases, as the previously utilized algorithms do not provide information about outliers. We consider a value of $MinPts = 3$ due to the small number of observations. In Figure 8, we present the graph of each country's distance from its 3rd closest neighbor. Through this graph, a suitable estimation of the parameter ε is 1500.

The countries corresponding to each cluster can be easily determined from the information provided by Figure 7. We observe the formation of four clusters and the marking of six irregular points. Focusing on the points marked as outliers, we can see that the algorithm characterized Denmark as an extreme case, a conclusion that is quite expected after the above comments on the country's energy policies. On the other hand, Norway, Iceland and Luxembourg are considered outliers due to their low kWh value compared to their high mean incomes.

This last observation explains that these 3 countries provide the most advantageous balance between income and electricity prices to their citizens. Quite positive is also the balance offered by Sweden and Finland—based on Figure 7—leading to the general conclusion that the northern countries display the most advantageous conditions for their population. Finally, countries such as Greece, Latvia, Portugal, Poland, Romania, Germany and Denmark seem to behave in a completely opposite way, as their electricity prices are relatively overpriced compared to the income of their population.

4. Discussion

In this study, we investigated the national kWh prices in relation to the corresponding average European household earnings using established clustering methods like K-means++, CLARA, PAM, UPGMA, and DBSCAN. Based on the clustering methodology, we distinguish cases of extreme overpricing and explore the external factors that cause these significant differentiations. At the same time, the proposed methodology reveals four price and salary groups that indicate the existence of geographical patterns from east to west and from south to north.

Regarding the examination of national electricity prices, our analysis shows an increasing trend in the kWh values. Countries such as Iceland, the United Kingdom, Estonia, Romania, Hungary, and Belgium are included in groups with higher kWh prices in 2018 compared to the period 2007–2017.

On the other hand, there are almost no countries included in clusters with overall lower kWh prices in 2018 compared to the 2007–2017 clustering results, confirming the phenomenon described above. This observation illustrates that the recent increase in energy prices should not be seen entirely as an unexpected event, but also as a situation that has emerged and gradually developed over the years.

The clustering results for both income and electricity prices are representative of each country over this period. This derives from the fact that each country was ranked in the cluster to which it belonged for the most years of that decade. The year 2018 was chosen for the comparisons with the clustering results for the ten-year period, due to the notable imbalances occurring this year.

Based on the clustering results for the mean family incomes of the 35 countries participating in the analysis, we observe a clear differentiation between the northwestern and southeastern European countries. The further we move from east to west and from south to north, the more satisfactory incomes become for the European population. In contrast, the respective national electricity prices do not follow this geographical pattern, leading to significant imbalances. In many countries, such as Greece, Poland, Romania, Slovakia, Croatia, Latvia, Hungary, the Czech Republic, Italy, Portugal and Spain, the kWh is overpriced compared to the mean incomes of an average household, with Portugal being the most representative example of this phenomenon. All the examples mentioned, are Eastern European countries or countries around the Mediterranean Sea.

On the other hand, the northern and Scandinavian countries offer the most favorable conditions. Representative instances of this phenomenon are Finland, France and the Netherlands, while this phenomenon is reinforced in countries like Iceland, Norway and Luxembourg. These three countries are simultaneously classified in the cluster with the most prosperous wages and in the cluster with the third most expensive kWh prices. Moreover, these three cases are identified as outliers by the DBSCAN algorithm (Figure 8), showing a deviant behavior compared to the other 32 European cases.

Finally, the clustering analysis has succeeded in separating the three northern countries with the most expensive kWh, whose final price is based on important exogenous factors. As we mentioned earlier, electricity prices in Germany and Denmark are characterized by increased taxation, corresponding to the effort to support sustainable energy sources. In addition, the high Belgian kWh value is due to disruptions in nuclear power plants, which led to the import of electricity from France. As a result, we note that in the only cases of the northern countries with relatively overpriced electricity, the final prices are significantly affected by additional parameters.

To sum up, after the abovementioned investigation we find an increasing trend in energy prices across Europe. Undoubtedly, there is a clear pattern in kWh prices depending on the geographical location of each country. If we move from the northern to the southern and eastern regions of Europe, we observe significant variations in the balance between incomes and kWh values, leading to the conclusion that these regions are often characterized by excessive electricity prices. In contrast, conditions in northern and Scandinavian countries are much more favorable for the population. Even in the few cases where we observed a relative increase in energy prices, this phenomenon is influenced by exogenous factors and its effect seems to be temporary. The fact that the results studied come not from one but from a total of five algorithms drastically increases their trustworthiness, as these methods suggest quite similar clustering, even if their functioning is fundamentally different.

5. Conclusions

In this article, we examine electricity price levels in 35 European countries relative to their respective average household incomes. For our analysis, we use five robust clustering methodologies, namely K-means++, PAM, CLARA, UPGMA and DBSCAN. Examining the results obtained from not just one, but a total of 5 clustering algorithms greatly increases their credibility, as these methods suggest similar groupings, even though their operation is fundamentally different.

There seems to be an increasing trend in national electricity prices, as in 2018, countries such as Iceland, the United Kingdom, Estonia, Romania, Hungary, and Belgium are classified in clusters that show a higher kWh value compared to the 2007–2017 period. Electricity prices in many southern and eastern countries seem to be overpriced in relation to the respective national household incomes, with Portugal being the most characteristic case of this phenomenon. On the other hand, countries like Norway and Iceland show the exactly opposite behavior.

A subsequent study concerning electricity prices later than 2018, could provide more information on the phenomenon under study, which could lead to a validation or modification of the presented patterns. Another perspective for future work would be to apply a similar methodology to examine the prices of other energy products, such as natural gas or heating oil prices in Europe. This attempt could complement the results of this article, while the comparisons between the evolution of electricity, natural gas and heating oil prices could provide valuable conclusions about the current and future state of energy prices.

In summary, existing kWh prices in recent years have led to remarkable imbalances among the European countries. These imbalances ultimately affect national economies, and this situation is expected to worsen, especially due to the recent energy crisis. This observation suggests that the recent rise in energy prices should not be seen as a completely unexpected event, but rather as a phenomenon that has become apparent over the years and is systematically intensifying.

Acknowledgments

We would like to thank Professor Georgios Afendras for his valuable contribution and guidance during the preparation of the present analysis. We also want to thank Dimitrios-Panagiotis Papageorgiou and Rafael Saligkaras for their contribution to this research. The authors contributed equally to preparation of the manuscript.

Conflicts of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Papageorgiou V (2021) Brain tumor detection based on features extracted and classified using a low-complexity neural network. *Trait du Signal* 38: 547–554. https://doi.org/10.18280/ts.380302

2. Papageorgiou VE, Zegkos T, Efthimiadis GK, et al. (2022) Analysis of digitalized ECG signals based on artificial intelligence and spectral analysis methods specialized in ARVC. *Int J Numer Meth Biomed Eng* 38: e3644. https://doi.org/10.1002/cnm.3644
3. Khanam M, Mahboob T, Imtiaz W, et al. (2015) A survey on unsupervised machine learning algorithms for automation, classification and maintenance. *Int J Comput Appl* 119: 34–39. https://doi.org/10.5120/21131-4058
4. Omran M, Engelbrecht A, Salman AA (2007) An overview of clustering methods. *Intell Data Anal* 11: 583–605. https://doi.org/10.3233/IDA-2007-11602
5. Fraley C, Raftery AE (2002) Model-Based clustering, discriminant analysis, and density estimation. *J Am Stat Assoc* 97: 611–631. https://doi.org/10.2307/3085676
6. Namratha M (2012) A comprehensive overview of clustering algorithms in pattern recognition. *IOSR J Comput Eng* 4: 23–30. https://doi.org/10.9790/0661-0462330
7. Matuszewska-Janica A, Żebrowska-Suchodolska D, Mazur-Dudzińska A (2021) The situation of households on the energy market in the european union: Consumption, prices, and renewable energy. *Energies* 14: 6364. https://doi.org/10.3390/en14196364
8. Gostkowski M, Rokicki T, Ochnio L, et al. (2021) A clustering analysis of energy consumption in the countries of the visegrad group. *Energies* 14: 5612. https://doi.org/10.3390/en14185612
9. Poyrazoglu G (2021) Determination of price zones during transition from uniform to zonal electricity market: A case study for Turkey. *Energies* 14: 1014. https://doi.org/10.3390/en14041014
10. Mart ńez Álvarez F, Troncoso A, Riquelme JC, et al. (2007) Discovering patterns in electricity price using clustering techniques. *International Conference on Renewable Energies and Power Quality—ICREPQ*, 174–181. https://doi.org/10.24084/repqj05.245
11. Verbič M, Filipović S, Radovanović M (2017) Electricity prices and energy intensity in Europe. *Util Policy* 47: 58–68. https://doi.org/10.1016/j.jup.2017.07.001
12. Gil-Alana LA, Martin-Valmayor M, Wanke P (2020) The relationship between energy consumption and prices. Evidence from futures and spot markets in Spain and Portugal. *Energy Strategy Rev* 31: 100522. https://doi.org/10.1016/j.esr.2020.100522
13. Takentsi S, Sibanda K, Hosu YS (2022) Energy prices and economic performance in South Africa: an ARDL bounds testing approach. *Cogent Econ Finance* 10: 2069905. https://doi.org/10.1080/23322039.2022.2069905
14. Shah I, Igtikhar H, Ali S (2022) Modeling and forecasting electricity demand and prices: A comparison of alternative approaches. *J Math* 2022: 3581037. https://doi.org/10.1155/2022/3581037
15. Wang J, Li H (2021) The impact of electricity price on power-generation structure: Evidence from China. *Front Environ Sci* 9: 733809. https://doi.org/10.3389/fenvs.2021.733809
16. Papageorgiou V (2022) A study of primary school teachers' tendencies regarding the usefulness of dramatization in the educational process. *Int J Cognitive Res Sci, Eng Educ (IJCRSEE)* 10: 145–162. https://doi.org/10.23947/2334-8496-2022-10-2-145-162
17. Zhou Y, Zheng S (2020) Machine-learning based hybrid demand-side controller for high-rise office buildings with high energy flexibilities. *Appl Energy* 262: 114416. https://doi.org/10.1016/j.apenergy.2019.114416

18. Zhou Y, Zheng S, Liu Z, et al. (2020) Passive and active phase change materials integrated building energy systems with advanced machine-learning based climate-adaptive designs, intelligent operations, uncertainty-based analysis and optimisations: A state-of-the-art review. *Renewable Sustainable Energy Rev* 130: 109889. https://doi.org/10.1016/j.rser.2020.109889
19. Liu J, Zhou Y, Yang H, et al. (2022) Uncertainty energy planning of net-zero energy communities with peer-to-peer energy trading and green vehicle storage considering climate changes by 2050 with machine learning methods. *Appl Energy* 321: 119394. https://doi.org/10.1016/j.apenergy.2022.119394
20. Day WHE, Edelsbrunner H (1984) Efficient algorithms for agglomerative hierarchical clustering methods. *J Classif* 1: 7–24. https://doi.org/10.1007/BF01890115
21. Murtagh F (1984) Complexities of hierarchic clustering algorithms: The state of the Art. *Comput Stat Q* 1: 101–113. Available from: https://www.researchgate.net/profile/Fionn-Murtagh-2/publication/238655641_Complexities_of_hierarchic_clustering_algorithms_State_of_the_art/links/5452a2970cf26d5090a377f1/Complexities-of-hierarchic-clustering-algorithms-State-of-the-art.pdf.
22. Murtagh F, Legendre P (2011) Ward's hierarchical clustering method: Clustering criterion and agglomerative algorithm. *J Classif* 31: 274–295. https://doi.org/10.1007/s00357-014-9161-z
23. Jain AK, Dubes RC (1988) Algorithms for clustering data. *Prentice-Hall, Inc. Division of Simon and Schuster One Lake Street Upper Saddle River, NJUnited States*. Available from: https://homepages.inf.ed.ac.uk/rbf/BOOKS/JAIN/Clustering_Jain_Dubes.pdf.
24. Caliński T, Harabasz J (1974) A dendrite method for cluster analysis. *Commun Stat* 3: 1–27. https://doi.org/10.1080/03610927408827101
25. Lance G, Williams WT (1967) A general theory of classificatory sorting strategies: 1. Hierarchical Systems. *Comput J* 9: 373–380. https://doi.org/10.1093/COMJNL/9.4.373
26. Lloyd SP (1957) Least squares quantization in PCM. *Technical Report RR-5497*, Bell Lab. Available from: http://mlsp.cs.cmu.edu/courses/fall2010/class14/lloyd.pdf.
27. Papageorgiou V, Tsaklidis G (2021) Modeling of premature mortality rates from chronic diseases in Europe, investigation of correlations, clustering and granger causality. *Commun Math Biol Neurosci*. https://doi.org/10.28919/cmbn/5926
28. Arthur D, Vassilvitskii S (2007) K-means++: The advantages of careful seeding. *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*. https://doi.org/10.1145/1283383.1283494
29. Hamerly G, Elkan C (2002) Alternatives to the k-means algorithm that find better clusterings. *CIKM '02: Proceedings of the Eleventh International Conference on Information and Knowledge Management*, 600–607. https://doi.org/10.1145/584792.584890
30. Kaufman L, Rousseeuw PJ (2005) Finding groups in data, an introduction to cluster analysis. *John Wiley Sons*. https://doi.org/10.1002/9780470316801
31. Ester M, Kriegel HP, Sander J, et al. (1996) A density-based algorithm for discovering clusters in large spatial databases with noise. *Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining.* Available from: https://www.aaai.org/Papers/KDD/1996/KDD96-037.pdf.
32. Saligkaras D, Papageorgiou VE (2022) Seeking the truth beyond the data. An unsupervised machine learning approach. https://doi.org/10.48550/arXiv.2207.06949

33. Kumar AV, Selvaraj JC (2016) A review on clustering algorithms. *J Recent Res Appl Stud* 8: 99–103. https://doi.org/10.2307/3085676
34. Gu J (2021) Comparative analysis based on clustering algorithms. *J Phys: Conference Series* 1994: 012024. https://doi.org/10.1088/1742-6596/1994/1/012024
35. Ankerst M, Breunig MM, Kriegel HP, et al. (1999) OPTICS: Ordering points to identify the clustering structure. *ACM SIGMOD Rec* 28: 49–60. https://doi.org/10.1145/304182.304187
36. Diaz-Rozo J, Bielza C, Larranaga P (2018) Clustering of data streams with dynamic gaussian mixture models: An IOT application in industrial processes. *IEEE Int Things J* 5: 3533–3547. https://doi.org/10.1109/JIOT.2018.2840129