energies

Analysis of the Behavior Pattern of Energy Consumption through Online Clustering Techniques

Juan Viera[1], Jose Aguilar [1,2,3,4,*], Maria R-Moreno[1,5] and Carlos Quintero Gull[6]

[1] Universidad de Alcalá, Escuela Politécnica Superior, ISG, Alcalá de Henares, 28805, Spain
[2] Universidad de Los Andes, CEMISID, Mérida, 5101, Venezuela
[3] Universidad EAFIT, GIDITIC, Medellín, 50022, Colombia
[4] IMDEA Networks Institute, Leganés, Madrid, Spain
[5] TNO, Intelligent Autonomous Systems Group (IAS), The Hague, The Netherlands
[6] Dpto de Ciencias Aplicadas y Humanísticas. Universidad de Los Andes, Mérida 5101, Venezuela
* Correspondence: Jose Aguilar

Abstract: Analyzing energy consumption is currently of great interest to define efficient energy management strategies. In particular, studying the evolution of the behavior of the consumption pattern can allow energy policies to be defined according to the time of year. In this sense, this work proposes to study the evolution of energy behavior patterns using online clustering techniques. In particular, the centroids of the groups constructed by the techniques will represent their consumption patterns. Specifically, two unsupervised online machine learning techniques ideal for the stated objective will be analyzed, X-Means and LAMDA, since they are capable of varying and adapting the number of clusters at runtime. These techniques are applied to energy consumption data in commercial buildings, making groupings on previous groups, in our case, monthly and quarterly. We compare their performance, to finish by analyzing the evolution of the patterns over time. The results are very promising since the quality of the consumption patterns obtained is very good according to the performance metrics. Thus, the three main contributions of this article are to propose an approach to determine energy consumption patterns using online non -supervised learning approaches, a methodology to analyze and explain the evolution of energy consumption using centroids of clusters, and a comparison strategy of online learning techniques. The online clustering techniques have qualities of the order of 0.59 and 0.41 for Silhouette and Davies-Boulding, respectively, for X-Means, and of the order of 0.71 and 0.24 for Silhouette and Davies-Boulding, respectively, for LAMDA, in different datasets of energy. The results are motivating, since very good results are obtained in terms of the quality of the clusters, particularly with LAMDA, therefore, analyzing its centroids as the patterns of user behaviors makes a lot of sense.

Keywords: Online clustering techniques; Energy consumption; LAMDA; X-means; Machine learning

1. Introduction

Currently, there is an immense global demand for energy, which is necessary for the functional consumption of most tasks of life, such as lighting, the use of computer equipment, household appliances, and other electronic devices. The aforementioned devices are currently vital in our society. On the other hand, currently, different types of buildings (residential, commercial, and industrial, among others) are being equipped with intelligent devices, such as cameras, sensors, and different actuators [1]. These devices, together with the communication infrastructure, characterize the Internet of Things (IoT) paradigm [2].

Due to the increase in energy required by this paradigm, consumption in homes has increased between 1,232 and 1,460 kWh per year [3]. It is estimated that the energy consumption derived from this increase in devices will increase much more in the coming years. For example, in Europe, it will go from 4 TWh in 2015 to 104 TWh in 2025 [3]. Due to this increase in energy demand, there is a great concern to achieve greater efficiency and optimization of consumption [1], [4], [5]. To do this, among other things, it is necessary to identify the consumption pattern of users, and based on that information, propose strategies and mechanisms to save energy resources as much as possible. In the next two sections, we compare other related works and describe the contributions of this paper. Particularly, being able to know how the customer's energy consumption pattern evolves can be useful in different energy management tasks [6]. For example, in the case of providers to determine when there are more or fewer demands to adjust their offer, and in the case of the client to know their peaks and from there look for an energy optimization mechanism.

1.1. Related work

Based on the interest of this article, which is to develop methods that allow knowing how the pattern of energy consumption of a client evolves, in this section, we will describe those recent works close to this topic. Particularly, since we have not found any work on this specific aspect, we present recent works linked to related topics, specifically, on the prediction and optimization of energy consumption.

In general, the current applications of artificial intelligence, and specifically machine learning algorithms, in the field of energy are enormous [6], [7], [8]. However, there are no recent works on the study of behavioral patterns of energy consumption. Most of the research works on the study of energy consumption and its characteristics focus on reducing consumption and optimizing the use of energy resources using, for example, optimization and predictive models. For example, in [8], Yoon et al. focus on the efficient use of energy and its infrastructure in smart cities using machine learning techniques. The authors use machine learning to create a deep learning network in a smart city to analyze and predict the energy consumption of IoT sensor devices. Wu et al. [10] use ML models to predict the consumption of an intelligent building, with the aim of energy conservation and environmental protection.

Xiao et al. [9] carried out a comparison of different configuration parameters for energy models. They propose 2 scenarios, one only with data to predict energy efficiency, and another that considers information from spindle motor aging and tool wear. For both cases, they use Support Vector Regression models [14], Artificial Neural Networks [15], and Gaussian Process Regression [16]. In the article [17], the authors proposed a predictive model based on the energy consumption of the users, which allows monitoring and estimating the energy consumption. In this case, they make use of the K-means algorithm and Support Vector Machines.

Other articles that analyze energy consumption, study strategies, and make predictions about energy consumption, among other things, are presented in [11] (based on Support Vector Regression, [12] (based on Artificial Neural Networks), and [13] (based on Random Forests). Also, for an analysis of user trends, there are currently various algorithms and methods, as well as techniques to reduce the complexity of the problem [9], [10], [18], [19], [20]. As we can see in this section, and to our knowledge, there are no previous works in the literature dedicated to studying the evolution of energy consumption patterns. Most of the works are dedicated to predict energetic behavior, to diagnose what may happen in an energetic infrastructure, but none of them have focused on defining energy consumption patterns and monitoring their evolution, and from there, proposing strategies of analysis and explainability of these patterns.

1.2. Contributions of the paper

The objective of this work is to identify and analyze the behavior pattern of customers according to their energy consumption profiles. In particular, it is necessary to identify how the pattern of customer behavior changes over time. We propose to use online unsupervised machine learning algorithms to follow/analyze the evolution of energy consumption patterns. For this, we assume that the centroids of the groups obtained by the clustering techniques represent the energy consumption patterns of the group. The main contributions of the work are:

- We propose a framework to analyze the evolution of energy consumption patterns.
- We adjust two clustering techniques to carry out an online clustering process of energy consumption data.

The work is organized as follows, section 2 presents the unsupervised machine learning used in this work. Section 3 describes the experiments and carries out an analysis of the clusters obtained. Section 4 presents an analysis of the evolution of the patterns and a general comparison in different datasets. Finally, the last section presents the conclusions and future work.

2. Online Unsupervised Machine Learning Used

Unsupervised learning algorithms assume that the data is not labeled and they analyze datasets to identify similarities between the data (similar data make up a cluster) [21]. This paradigm is useful when the categories of the data are not defined, and one of the techniques used in this area is clustering algorithms [22]. The main purpose of a clustering algorithm is to separate the data into smaller subsets, called groups (clusters), such that the content of the data is similar in each cluster but different from the content of the other groups. The centroid of a cluster can be understood as its pattern. Particularly, we will use unsupervised online learning to adapt the cluster to changes in user consumption patterns, enabling real-time updates [23]. In this work, we will use X-means and the LAMDA algorithms.

2.1. X-Means

The X-means algorithm is based on K-Means. The K-means algorithm is one of the simplest and most common algorithms used in clustering, dividing the dataset into K clusters. K-means tries to find the center of each cluster, which is representative of a data region [21]. This point is called the centroid. Thus, K-means is a clustering technique based on centroids [22]. K-means alternates 2 steps:

- Assignment of points/individuals to the nearest centroid
- Calculation of centroids

These steps are repeated in a loop until the centroids stabilize.

Particularly, in this work, we will use X-Means, which is an extension of K-Means that allows varying the value of K (it does not have to be predefined at the beginning, as it happens with K-Means) [24]. Thus, X-means is an incremental sequential K-means that determines the value of K (clusters) based on a function *f(K)*, which is defined by the following Equation [25]:

$$f(K) = \begin{cases} 1 & if K = 1 \\ \dfrac{S_K}{\alpha_K S_{K-1}} & if S_{k-1} \neq 0, \forall K > 1 \\ 1 & if S_{k-1} = 0, \forall K > 1 \end{cases} \tag{1}$$

Where $$\alpha_K = \begin{cases} 1 - \frac{3}{4N_d} & if K = 2 \wedge N_d > 1 \\ \alpha_{K-1} + \frac{1-\alpha_{K-1}}{6} & if K > 2 \wedge N_d > 1 \end{cases}$$

Where S_k is the sum of the cluster distortions when the number of clusters is K (see below), and N_d is the number of attributes in the dataset. The term $\alpha_k S_{k-1}$ in the Equation above is an estimate of S_k based on S_{k-1}, made under the assumption that the data have a uniform distribution. The value of *f(K)* is the ratio of the actual distortion to the estimated distortion, and is close to 1 when the data distribution is uniform. When there are areas of concentration in the data distribution, then S_k will be less than the estimated value, so *f(K)* decreases. The *smaller f(K),* the more concentrated the data distribution. Therefore, values of K that produce a small value of *f(K)* can be considered to provide well-defined groups.

On the other hand, the distortion of a cluster is the distance between the objects/individuals of a cluster and its centroid, according to the following Equation [25]:

$$I_j = \sum_{t=1}^{N_j} \left[d\left(x_{jt}, w_j\right)\right]^2 \quad (2)$$

Where I_j is the distortion of cluster j, w_j is the centroid of cluster j, N_j is the number of objects belonging to cluster j, x_{jt} is the object t belonging to cluster j, and $\mathrm{d}(x_{jt}, w_j)$ is the distance between the object x_{jt} and the centroid w_j of cluster j. Each cluster is represented by its distortion, and the overall impact of all clusters on the entire data set is evaluated by the sum of all distortions, S_K, given by the following Equation [25]:

$$S_K = \sum_{j=1}^{K} I_j \quad (3)$$

Where K is the number of clusters. The number of clusters K is assumed to be much smaller than the number of objects N. In particular, if for any immediate K *f(K)* shows special behavior, in particular a minimum point, that value of K should be taken as the number desired of clusters. Thus, X-Means converges when it obtains a minimum value of *f(K).*

In this way, X-means determines if new centroids should appear within a current model (M_j). The appearance of new centroids is carried out by dividing some clusters into 2, which have been classified as optimizable according to the Schwarz criterion (it is a criterion for the selection of models among a finite set of models), based on the BIC value, defined by the following equation [24]:

$$BIC\left(M_j\right) = \hat{l}_j(D) - \frac{p_j}{2}.logR \quad (4)$$

Where, $\hat{l}_j(D)$ is the logarithmic probability of the data in the model M_j; p_j is the number of free parameters present in the model M_j; and R represents the number of samples present in *D* (*R* = |*D*|).

In essence, X-means starts with a given K, goes on to add centroids (changes the K) according to the value of *f(K),* and calculates the BIC score for each cluster to determine, if any, which cluster to split. When X-Means converges (determines the ideal value of K for that data set), then the final clustering is obtained.

2.2. LAMDA (Learning Algorithm for Multivariate Data Analysis)

LAMDA is a non-iterative fuzzy algorithm based on the degree of adequacy of an individual (data) to a group. It provides great versatility since it allows not to specify the number of clusters during the execution and, furthermore, it can work online [26], [27]. LAMDA works by performing an evaluation of the similarity between the descriptors of an element X of the form X = $\{x_1, x_2, \ldots, x_j, \ldots x_m\}$, which is its vector with m descriptors, with the descriptors of the centroids of the existing clusters, to define in which cluster this data X should be entered. In addition, once X has been assigned to a cluster, it becomes X

= $\{x_1, x_2, \ldots, x_j, \ldots x_m, c_i\}$, i = 1, 2, … k, where c_i is the label associated with X [26]. The base definitions of LAMDA are summarized below [26], [28].

Normalization. Each descriptor of X must be normalized, based on its maximum and minimum values:

$$\overline{x_j} = \frac{x_j - x_{jmin}}{x_{jmax} - x_{jmin}} \tag{5}$$

Where ($\overline{x_j}$) is the normalized value of descriptor j, x_{jmin} is the minimum value of descriptor j, and x_{jmax} is the maximum descriptor of descriptor j. The element resulting from normalization X will be used to compute the degree of adequacy of the element to each existing cluster.

Degree of Marginal Adequacy (MAD). Determines the degree of similarity of a descriptor with respect to another descriptor in a given class. For the calculation of the MAD, density functions are used, the most common is the fuzzy binomial function:

$$MAD(\overline{x_j}/\rho_{kj}) = \rho_{kj}^{\overline{x_j}} (1 - \rho_{kj})^{(1-\overline{x_j})} \tag{6}$$

Where ρ_{kj} is the mean value of descriptor j in the cluster k, calculated by:

$$\rho_{kj} = \frac{1}{n_{kj}} \sum_{t=1}^{n_{kj}} \overline{x_j}\ (t) \tag{7}$$

ρ_{kj} is progressively updated each time a new element is added to the cluster.

The function for $MAD(\overline{x_j}/\rho_{kj})$ is the density function of the binomial distribution, which can be interpreted as the probability that the analyzed normalized descriptor belongs to a cluster j, given its mean ρ_{kj}.

Degree of Global Adequacy (GAD). Determines the degree of adequacy of a sample to each existing cluster, it is calculated by mixing the MAD with aggregation functions. These functions are interpolations between the t-norm (T) and the t-conorm (S), like the Dombi operator [29]:

$$T(a,b) = \frac{1}{1 + \sqrt[p]{(\frac{1-a}{a})^p + (\frac{1-b}{b})^p}} \tag{8}$$

$$S(a,b) = 1 - \frac{1}{1 + \sqrt[p]{(\frac{1-a}{a})^p + (\frac{1-b}{b})^p}} \tag{9}$$

In most cases, p = 1 is used to obtain an approximation close to a linear behavior of the t-norm and the t-conorm [29].

There is also a requirement parameter 0 <α < 1, used to calibrate fuzzy partitioning data [30]. If α = 1 then GAD is calculated as the t-norm, obtaining a stricter clustering. If α = 0 then GAD is computed as a t-conorm, leading to a more permissive grouping. Thus, α produces a linear interpolation between the t-norm and the t-conorm to calculate the GAD [31].

$$GAD_{\bar{X},k}(MAD_{k,1}, \ldots, MAD_{k,1}) = \alpha\, T(MAD_{k,1}, \ldots, MAD_{k,1}) + (1-\alpha)\, S(MAD_{k,1}, \ldots, MAD_{k,1}) \tag{10}$$

On the other hand, when an individual (data) does not belong to any class, then a non-informative class (NIC) is created, which will be a new cluster. The GAD of the data entering the NIC is computed considering that $MAD_{NIC\ j} = 0.5$, independent of the value of $\overline{x_j}$:

$$GAD_{\bar{X},NIC} = \alpha\, T(0.5,\ldots,0.5) + (1-\alpha)\, S(0.5,\ldots,0.5)$$

That element that enters the NIC becomes the first element of the new cluster.

Finally, the assignment of elements to a cluster is done by calculating the maximum GAD of all classes. The index (in) corresponds to the number of the class where the element will be assigned:

$$\text{in} = \max\,(GAD_{1\,\bar{X}}, GAD_{k\,\bar{X}}, \ldots, GAD_{m\,\bar{X}}, GAD_{NIC\,\bar{X}})$$

3. Experiments

In this section, we will explain how we perform the instantiation and execution of the two techniques presented in the previous section.

3.1. Data Preparation

For this experiment, we have used a real dataset from [32]. The first task is to divide the dataset into several files by time periods. In our case, they were divided by months or quarters. From the original data, more data has been generated using the distribution of each variable in the dataset, in order to increase the amount of data for our execution.

This first dataset corresponds to data taken from a commercial building in 2018. The building had a maximum hourly consumption of 48 W/m^2, and the annual consumption was 183.2 kWh/m^2 [32]. Each variable in the dataset was taken every half hour throughout the year, breaking down the total consumption in kW as follows: total consumption, light, heat pump, air treatment units, circulation pumps, heating and hot water, cooling, air coolers, and elevators.

3.2. Metrics

To evaluate the quality of the online clustering algorithms we have used two metrics. An ideal metric for distance-based algorithms like X Means (Silhouette coefficient [33]), and another metric for density-based algorithms (Davies-Bouldin index [34]).

Silhouette

The Silhouette coefficient is a measure of the cohesion of the clusters. It determines the degree of similarity between the objects of the same cluster [33]. To get this measurement, the average of the proximities between its elements is calculated. This metric is therefore effective in situations where the clusters have a circular shape [23], [33] or are grouped around a point. The silhouette coefficient for a data sample is determined with the mean of the silhouette coefficient for each sample data, calculated as [33]:

$$S_S = \sum_{i=1}^{n} \frac{b(i)-a(i)}{\max\{a(i), b(i)\}} \quad (11)$$

Where a(i) and b(i) are computed for each sample i of the cluster C_i $(i \in C_i)$ for $a(i) = (|C_i| - 1)^{-1} \sum_{j\in C_i,\ i\neq j} d(i,j)$ and $b(i) = min_{k\neq i} |C_k|^{-1} \sum_{j\in C_k} d(i,j)$, where d(i, j) is the distance between the points i and j.

The coefficient gives a result between -1 and 1. Values close to 1 are the most optimal, those close to 0 indicate that there are overlapping clusters, and negative values generally indicate that there are samples erroneously assigned to clusters. As a general rule, the higher the silhouette coefficient, the better defined the clusters will be [23].

Davies-Bouldin

The Davies-Bouldin index is defined as the mean similarity of each cluster with its most similar cluster. This measure compares the distance between both clusters with the size of the clusters themselves [34]. The measure can be used to infer the adequacy of a data partition. The Davies-Bouldin index is calculated as [34]:

$$S_{DB} = \frac{1}{k}\sum_{i=1}^{k} max_{i \neq j}\, R_{ij} \tag{12}$$

Where R_{ij} is the similarity between the clusters *i* and *j*. There are different ways to calculate R_{ij}, one of them is $R_{ij} = {s_i + s_j}/{d_{ij}}$, where s_i is the average distance between each point of cluster i and the center of cluster i, and d(i, j) is the distance between the centroids of the clusters i and j.

The minimum value that can be obtained using this index is 0, which is the case when there are as many clusters as there are individuals. Therefore, it is understood that the best values of this metric are those closest to 0 since they indicate a better partition and a model with better separation between clusters [23], [34].

3.4. Modeling

Next, we proceed to describe how the clustering models are obtained with each algorithm, using a time period (iteration) of a month.

X-means

In the first iteration, k is initialized to 3 (number of initial clusters), a value that X-Means then optimizes in that first iteration (month). In the following iterations (months), the algorithm readjusts that value of K. In the specific case of the dataset used, X-Means determines that 20 clusters are necessary on its second iteration. This number of clusters is maintained throughout the 12 months, X-Means determines that it is the ideal value of K in each iteration (month).

We will start by evaluating the centroids of the 20 clusters from January to December in Figures 1 and 2. The centroids for the analysis are normalized between 0 and 1 to graph them (it is the X axis of Figures 1-5), and then the energy consumption represented by them is what we analyze next. Looking at both Figures, it can be seen that in a range of approximately 0.14 and 0.06 in the centroids (equivalent to a range between 200kW and 400kW over the total of kW), there are 10 clusters. We also see 2 clusters in the upper range of Figure 2, which stand out for being separated from those in the middle zone. Clusters 19 and 20 are isolated from the rest throughout the run, slightly converging and stabilizing at the end of the run. Particularly, in Figure 2, we see that in the summer months, clusters 19 and 20 behave erratically, perhaps with a greater number of clusters this behavior would be softened.

In this particular case, clusters 19 and 20 represent a high consumption, in one case the consumption is higher due to the circulation pumps, and in the other case due to the heat pump and heating and hot water. Finally, cluster 5 represents the pattern with the lowest consumption (around 250kW), which is generated by several variables (lights, refrigeration, and elevators).

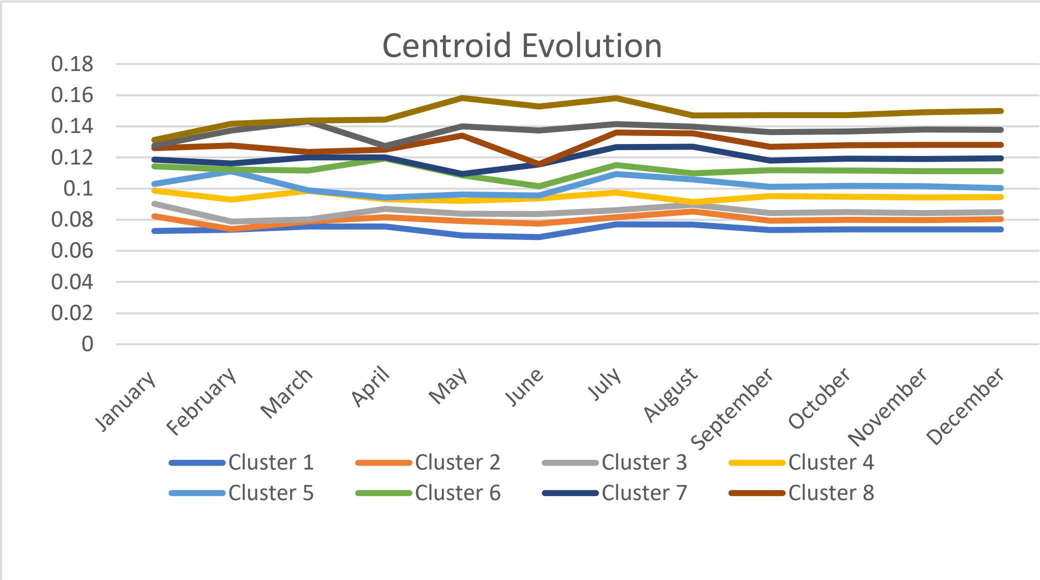

Figure 1. Evolution of the centroids of the first 10 groups with X-means.

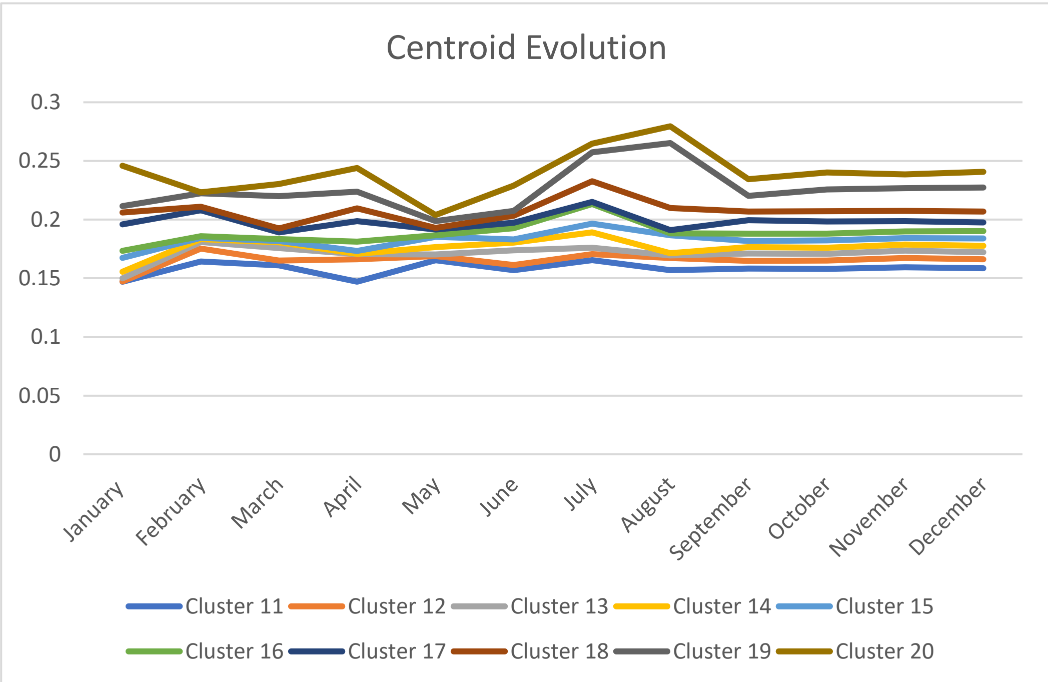

Figure 2. Evolution of the centroids of the last 10 with X-means.

Limiting the upper range of clusters to 15 (this value is when X-Means has the best performance), we obtain a more detailed view of the clusters in Figure 3. We find 10 clusters that never exceed 0.15 (500kW of the total), regardless of the time of the year. On the other hand, we see in Figure 3 how in the last quarter the variations are minimal. It can be deduced from this that a suitable and stable grouping has been reached, with well-defined groups. Some clusters represent the consumption of more than 600 kW, such as clusters 14 and 15. According to their centroids, in one case it is for heat and circulation pumps, and in the other for air treatment units, cooling, and air coolers.

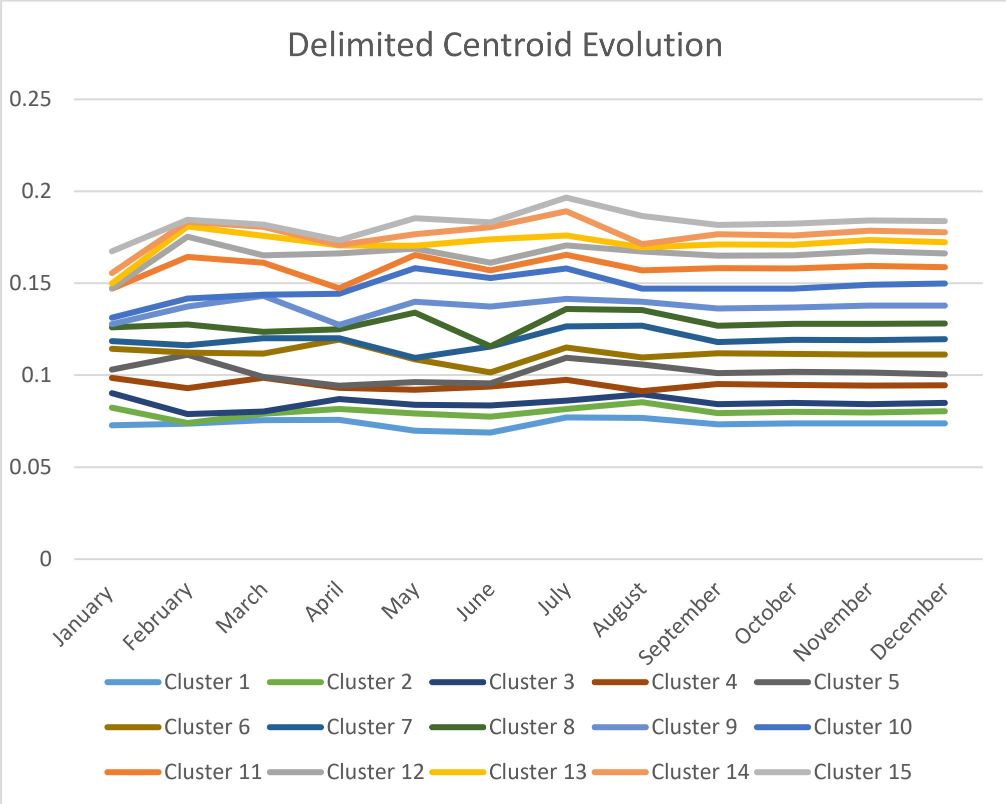

Figure 3. Bounded clustering with X-Means (K_{max}=15).

LAMDA

For the execution of LAMDA, an implementation of this algorithm has been used following what is indicated in the article [28]. In the same way as in the execution of X-means, the data is evaluated month by month. In the first iteration, the algorithm starts with a single empty cluster, and new clusters are created each time an element enters the NIC. We remember that the values that enter the NIC are those that have not managed to be located in existing clusters. All values are normalized before starting their evaluation. Particularly, the centroids are normalized between 0 and 1. LAMDA eliminates, merges and creates clusters depending on the GAD and the defined neighborhood threshold.

In Figure 4 can be seen how at the beginning of the execution, in January, 16 clusters are created, although 3 of them (10, 11 and 13) are merged after the first month. These remaining 13 clusters are maintained throughout the rest of the run. All the clusters arrive at a different point, except for the sets {7, 8} and {4, 5} whose centroids end up being quite similar, although their trajectory over the months is very different. In this case, the value of the centroids of clusters 10, 11 and 13 differ mainly in the variables light, air coolers, and elevators. Similarly, the difference among clusters 7 and 8 is mainly in the values of heat and circulation pumps, and in the case of clusters 4 and 5 of air treatment units, cooling, and air cooler.

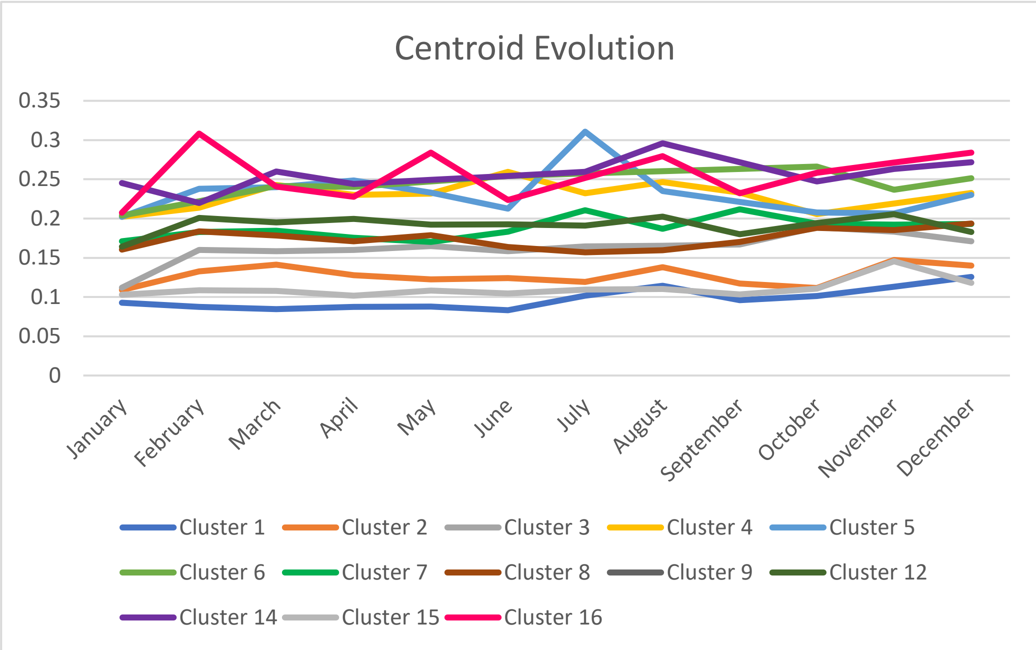

Figure 4. Evolution of the centroid of the first groups with LAMDA.

In Figure 5 we see the rest of the clusters created throughout the execution. Starting in February, new groups are being created, and it can be seen that there are stable clusters and others that vary over time. For example, cluster 33 completely changes its trend, going from being in a range of 0.2-0.25 in July to dropping to 0.09 in October, establishing itself as the only cluster in that low value. Here we can also see the last cluster that is created in August, this being number 40. Particularly, cluster 33 represents a decrease in energy consumption to less than 400 kW, derived mainly from the values of heat pump, heating and hot water, and cooling.

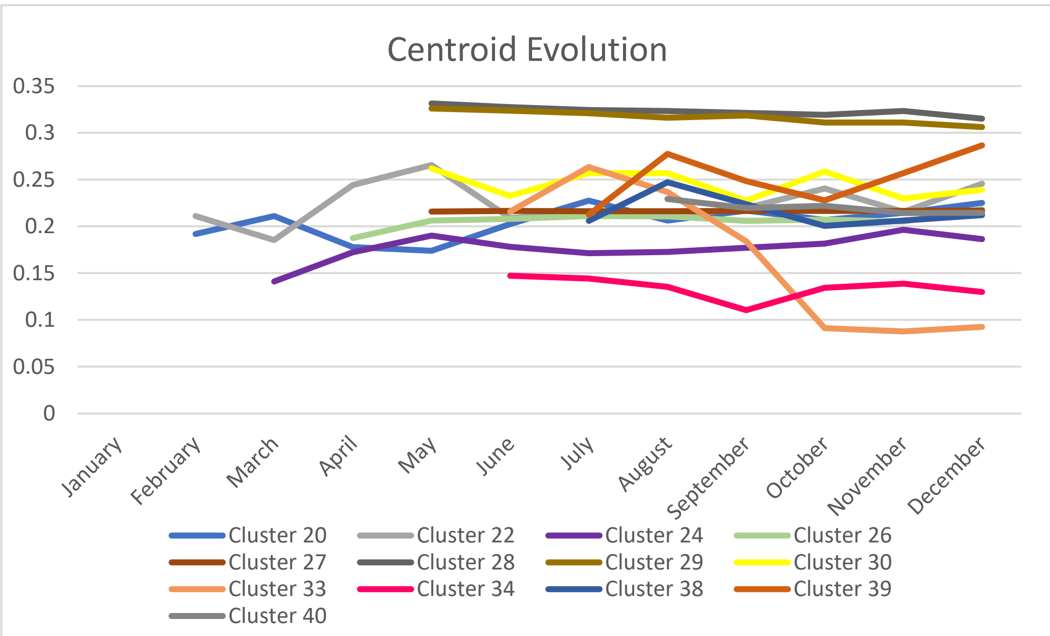

Figure 5. Evolution of the centroid of the last groups with LAMDA.

3.5. Comparison of both algorithms

Finally, Table 1 presents a comparison of the results of both algorithms.

Table 1. Results of the clustering algorithms.

	X-means		LAMDA	
	Silhouette	**Davies-Boulding**	**Silhouette**	**Davies-Boulding**
January	0,446	0,620	0,694	0,305
February	0,388	0,645	0,521	0,396
March	0,389	0,604	0,514	0,278
April	0,384	0,614	0,541	0,238
May	0,346	0,589	0,563	0,217
June	0,390	0,598	0,513	0,233
July	0,387	0,626	0,561	0,321
August	0,382	0,614	0,591	0,348
September	0,377	0,597	0,515	0,423
October	0,386	0,603	0,528	0,248
November	0,384	0,638	0,519	0,321
December	0,381	0,599	0,516	0,294

Based on the metrics, LAMDA consistently performs better on both metrics. On the other hand, the Silhouette coefficient is an excellent metric in data with circular spatial behavior, while Davies-Bouldin is better in other cases. According to the results obtained, it could be intuited that the spatial distribution of the data is circular, so silhouette would be the best metric to compare them. Now, X-Means does not change the clusters, while LAMDA adjusts the number of clusters over time. Thus, the advantage of LAMDA is that it automatically checks the need to merge and create new clusters. We are interested in studying this evolution in the next section.

4. Analysis of the evolution of clusters

In this section, we analyse the evolution of LAMDA clusters by month and quarterly. Subsection one studies in detail how LAMDA is creating and merging the clusters over time, and subsection 2 extends the periods to quarters, to evaluate the capacity of LAMDA for larger periods. Also, at the end, we discussed the size of the clusters.

4.1. Initial experiment

Comparing the evolution of both algorithms, at the end of the execution, there are 20 and 26 clusters for X-means and LAMDA, respectively. In this section, we will analyze the evolution of LAMDA clusters, since it presents the best results and has a more dynamic behavior, creating and merging clusters throughout the execution.

We will start by analyzing the creation and merger of clusters shown in Table 2. Let us remember that the online clustering process is cumulative, that is, the behavior of the previous month is taken into account. Table 3 shows the reference month, the identifier of the clusters formed, the total number of clusters formed at the moment, and also comments where it is mentioned if there is a merger of some clusters, as well as the number of clusters that are added in the month.

Initially, 16 clusters are formed, of which the clusters identified with the numbers 10, 11 and 13 merge with other clusters, leaving a total of 13 in the first month. For the second month, it is observed that apart from the 13 clusters created the previous month, initially 6 new clusters are added, of which the clusters with id 8, 17, 19 and 21 are merged, leaving a total of 15 clusters. The value of the centroids of clusters 10, 11 and 13 differ only in the variables of light, air coolers, and elevators. Similarly, the value of the centroids of clusters 8, 17, 19 and 21 differ only in the variables of air treatment units and air coolers.

Continuing with the analysis, the first 9 clusters generated in January are maintained throughout the study period. This behavior of creating and merging clusters is maintained until September, generating up to 40 clusters, of which 26 remain. As of September, no new clusters are created nor are there new mergers, such that all new observations/individuals are added to one of the 26 clusters formed so far.

Table 2. Creating and merging clusters with LAMDA.

Month	Id of clusters created	Total of clusters	Comments
1	1, 2, 3, 4, 5, 6, 7, 8, 9, 12, 14, 15, 16	13	16 clusters formed and the next clusters are merged: 10, 11 and 13
2	1, 2, 3, 4, 5, 6, 7, 8, 9, 12, 14, 15, 16, 20, 22	15	6 additional clusters are formed and the next clusters are merged: 17, 18, 19 and 21
3	1, 2, 3, 4, 5, 6, 7, 8, 9, 12, 14, 15, 16, 20, 22, 24	16	2 additional clusters are formed and the next cluster is generated: 23
4	1, 2, 3, 4, 5, 6, 7, 8, 9, 12, 14, 15, 16, 20, 22, 24, 26	17	2 additional clusters are formed and the next cluster is generated: 25
5	1, 2, 3, 4, 5, 6, 7, 8, 9, 12, 14, 15, 16, 20, 22, 24, 26, 27, 28, 29, 30	21	4 additional clusters are formed and there is no fusion
6	1, 2, 3, 4, 5, 6, 7, 8, 9, 12, 14, 15, 16, 20, 22, 24, 26, 27,28, 29, 30, 33, 34	23	Form 4 additional clusters and merge 31 and 32
7	1, 2, 3, 4, 5, 6, 7, 8, 9, 12, 14, 15, 16, 20, 22, 24, 26, 27, 28, 29, 30, 33, 34, 38, 39	25	Form 5 additional clusters and merge: 35 36 and 37
8	1, 2, 3, 4, 5, 6, 7, 8, 9, 12, 14, 15, 16, 20, 22, 24, 26, 27, 28, 29, 30, 33, 34, 38, 39, 40	26	1 additional cluster is formed and there is no fusion
9	1, 2, 3, 4, 5, 6, 7, 8, 9, 12, 14, 15, 16, 20, 22, 24, 26, 27, 28, 29, 30, 33, 34, 38, 39, 40	26	No additional cluster formation
10	1, 2, 3, 4, 5, 6, 7, 8, 9, 12, 14, 15, 16, 20, 22, 24, 26, 27, 28, 29, 30, 33, 34, 38, 39, 40	26	No additional cluster formation
11	1, 2, 3, 4, 5, 6, 7, 8, 9, 12, 14, 15, 16, 20, 22, 24, 26, 27, 28, 29, 30, 33, 34, 38, 39, 40	26	No additional cluster formation
12	1, 2, 3, 4, 5, 6, 7, 8, 9, 12, 14, 15, 16, 20, 22, 24, 26, 27, 28, 29, 30, 33, 34, 38, 39, 40	26	No additional cluster formation

4.2. Quarterly Evolution Analysis

We decide to analyze the evolution of the clusters by quarter. We can see in Figure 6 how the general tendency is to remain stable and follow a predictable trend. The especially erratic behavior that appeared in clusters 19 and 20 in Figure 2 is no longer visible. These clusters have few individuals compared to the rest of the groups, which makes them more volatile to small changes or new inclusions in the cluster. These are clusters that represent patterns with high consumption (more than 700 kW). Cluster 39 has a behavior pattern similar to that of 16, so, over time, if they maintain this trend, it is possible that they will unify because the difference is due to the values of the light and elevators. In the same way, we can study the behavior of 3 clusters that are approaching in December, these are clusters 26, 27 and 40, which are grouped below the 0.22 value. However, this case is different from the previous one since they only approach the end of the analysis, as we can see in Figure 10 (the difference is due to the values of air treatment units and air coolers). In this case, we should be aware of their evolution since the 3 come with different trajectories.

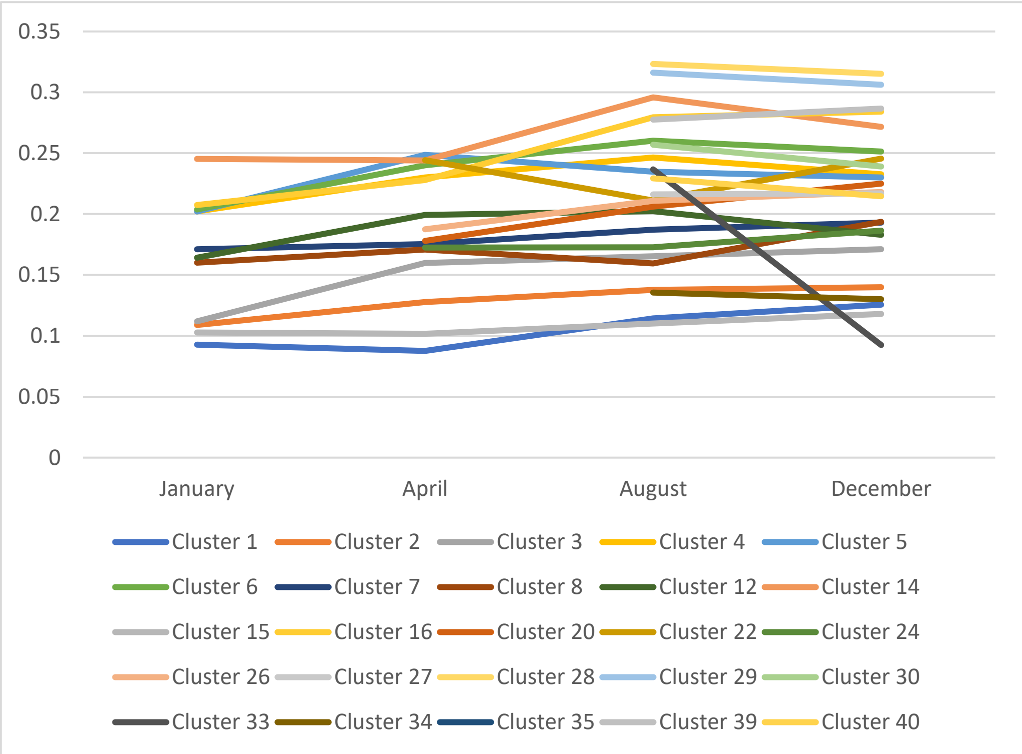

Figure 6. Evolution of the clusters by quarter with LAMDA.

Finally, Figure 7 shows the number of average individuals per cluster through the evolution of the clusters for this period of time. The most populated clusters are 15, 33 and 34. Groups 15 and 34 are quite populated throughout their existence, and 33 goes from being a cluster with few individuals to one with a lot of weight. This change occurs when, being in the range of 0.24 in August, it falls to 0.09, where it stabilizes. In these clusters, it can be seen that their trajectory (once they have a high number of elements) is more stable, and only small corrections are made to their centroids as individuals are added to the groups. We see then that the majority of individuals are in these 3 groups. Particularly, in December its centroids are 15 = 0.170, 33 = 0.092 and 34 = 0.119 (see Figure 6). Cluster 15 represents a pattern with medium consumption (more than 500 kW) due to mainly heat and circulation pumps, and heating and hot water. Similarly, cluster 34 represents a pattern with medium consumption (less than 400 kW) due to mainly cooling, air coolers and air treatment units. Finally, cluster 33 is a pattern of low consumption (less than 300 kW).

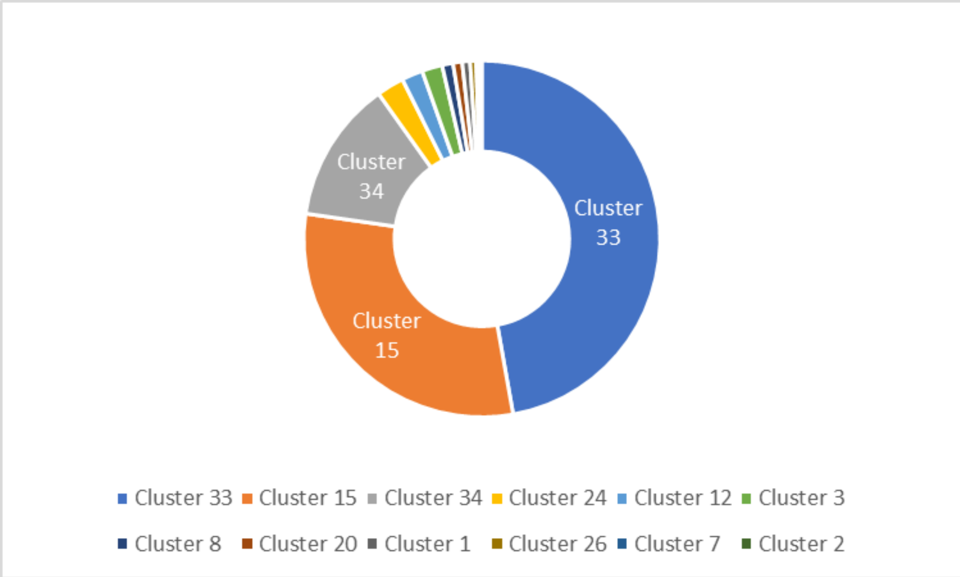

Figure 7. Partial view of the distribution of elements by cluster.

In Figure 7, we see how most of the elements/individuals have been assigned in these 3 clusters. Between them, they occupy almost 90% of the data occupation, the rest of the data is found in the remaining 23 clusters. With this, we can see that most of the elements are in the medium and low threshold of consumption, the centroid with the highest value of this trio of clusters is that of cluster 15, with 0.17 (less than 600 kW).

Let's analyze cluster 33, which has more individuals. In a cluster whose trend (evolution) reflects a fall in the last quarter of the year, which may be due to the fact that the use of these office/laboratory spaces is reduced at that time of the year. We can see the next values in the centroid variables, the average consumption of light of 3.5 kW, of the heat pump of 24.2 kW, of air treatment units of 2.8 kW, and circulation pumps of 0.42 kW. They reflect a space with a moderate consumption of energy, mainly derived from the consumption of the heat pump. As this is the device with the highest consumption, its use in heating and cooling tasks could be analyzed to optimize it.

As can be seen, the detailed analysis, both at the temporal level and at the level of the values of a pattern, allows two things: i) determine the energy behavior over time to establish temporary improvement measures (for example, in the months of greatest consumption search for less expensive energy sources) ii) determine the devices that consume the most, the reason, in order to establish strategies that optimize them.

5. Comparison in different datasets

To show the feasibility of the energy consumption evolution analysis process based on our online clustering algorithms, several energy consumption datasets are used in this section. Table 3 shows in the first column where the datasets were drawn from, and in the following columns the quality of the techniques in each of the performance measures analyzed in the work. This allows determining if the clusters obtained in each case are of high quality.

Table 3. Results of the clustering models in different datasets.

Dataset	Algorithm	Davies-Boulding	Silhouette
[35]	X-Means	0.349	0.893
	LAMDA	0.144	0.892
[36]	X-Means	0.331	0.575
	LAMDA	0.251	0.660
[37, 38]	X-Means	0.395	0.530
	LAMDA	0.257	0.633
[39]	X-Means	0.645	0.635
	LAMDA	0.291	0.690
[40]	X-Means	0.415	0.625
	LAMDA	0.177	0.669

According to the results, we see that LAMDA is a very robust method. In particular, in the different datasets, it obtains the best result. It is a very robust algorithm regardless of the energy consumption dataset (time series type). In addition, we see in the previous results (see section 5) the ability of LAMDA to create or merge clusters over time to adapt to the context.

6. Conclusion

In this work, we have performed online clustering algorithms to analyze the evolutionary behavior of energy consumption patterns, understood as the centroids of the groups they propose. By using X-means and LAMDA, we are able to delegate decision-making about the number of clusters to the algorithms. This was particularly shown in LAMDA since it was able to increase and/or decrease the number of groups. In X-Means, we couldn't see this behavior, since from the first iteration it created the maximum number of clusters. On the other hand, an analysis without a cluster limit is more appropriate in a real scenario (for example, in the case of X-means, the values of K were bounded in one case), regardless of the time it takes. In addition, with X-Means the abnormal behavior of some clusters was observed (affected by outliers).

The analysis of the centroids with LAMDA has made clear the great difference in consumption between users. In addition, according to its evolution, consumption trends can be studied. In short, the analysis of the evolution of the centroids of the groups allows making more precise decisions in the energy world (months of higher consumption, abnormal behavior, etc.). Thus, something relevant is how the variables that generate more energy consumption can be analyzed, particularly, the evolution of this consumption through the months (for example, cluster 33 in Figure 6). In general, in the patterns that represent high energy consumption, the variables responsible for this high energy consumption are clearly identified. Normally, these variables, in some cases were heat and circulation pumps, and heating and hot water, and in others were air treatment units, cooling, and air coolers. These combinations of variables are closely linked to high consumption. Likewise, there is a relationship between these variables with respect to the time of year, due to the environmental impact of the time of year on these variables. On the other hand, it can also be identified in the centroids that the variables that have very little impact on energy consumption are light and elevators.

Thus, we have shown in this work the feasibility of using online unsupervised learning approaches to monitor energy consumption patterns. In addition, with our approach, it is possible to analyze and explain in detail the evolution of energy consumption using the cluster centroids, with which it is possible to study their behavior over time, and determine the specific energy behavior of the devices. With both, optimization strategies can be defined, both at a global level (according to the customer's consumption trend) and at a specific level (in the devices).

In general, the pattern of energy consumption behavior of a customer/user can be used by both suppliers and consumers. In the case of consumers, know their energy consumption and, based on this, optimize it, carry out optimal management of it, among other things, and in the case of suppliers to adapt their offer to the needs of users, among other things. One of the limitations of this work is that it has been carried out with datasets, but in a real context, a robust platform will be required that captures in real time the different energy consumption values of the different devices to be monitored. Another limitation is the dependence on the quality of the data from the clustering process, which may affect the quality of the results when there are many atypical values, missing data, among other aspects.

Some aspects to take into account for future work are: i) Have data on energy consumption (applied in our case) together with user profile, to favor a more complete and specific analysis (for example, profiling the energy behavior of an individual) and; ii) an automatic construction of the analysis of the evolution of the clusters would be ideal (give more explainability to the centroids that are obtained), to help decision-makers. Therefore, a future work should define hybrid models that combine online clustering algorithms with techniques that allow predicting some of the energy variables. Also, this work will be extended to analyze these patterns using explainability techniques, to establish an interpretability of the patterns from the behavior of the attributes that make up the centroids. Finally, future works will use these results in an intelligent energy management system, in order to personalize their behavior in the function of the consumer's energy pattern.

Funding: J. Aguilar is supported by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 754382 GOT ENERGY TALENT. M.D. R-Moreno is supported by the JCLM project SBPLY/19/180501/000024 and the Spanish Ministry of Science and Innovation project PID2019-109891RB-I00, both under the European Regional Development Fund (FEDER).

References

[1] K. Akkaya, I. Guvenc, R. Aygun, N. Pala and A. Kadri, "IoT-based occupancy monitoring techniques for energy-efficient smart buildings," *201*, "IoT-based occupancy monitoring techniques for energy-efficient smart buildings," in - *IEEE Wireless Communications and Networking Conference Workshops (WCNCW),* 2015. DOI: 10.1109/WCNCW.2015.7122529.

[2] K. Patel, S. Patel, "Internet of Things-IOT: Definition, Characteristics, Architecture, Enabling Technologies", Application & Future Challenges. *International Journal of Engineering Science and Computing*, vol. 6, (5), pp. 6122-6131, 2016.

[3] C. Gray, R. Ayre, K. Hinton and L. Campbell, "'Smart' Is Not Free: Energy Consumption of Consumer Home Automation Systems," *IEEE Transactions on Consumer Electronics,* vol. 66, (1), pp. 87-95, 2020. DOI: 10.1109/TCE.2019.2962605.

[4] R. Yang and L. Wang, "Development of multi-agent system for building energy and comfort management based on occupant behaviors," *Energy Build.*, vol. 56, pp. 1-7, 2013. DOI: 10.1016/j.enbuild.2012.10.025.

[5] E. Fotopoulou, A. Zafeiropoulos, F. Terroso-Sáenz, U. Şimşek, A. González-Vidal, G. Tsiolis, P. Gouvas, P. Liapis, A. Fensel, A. Skarmeta,, "Providing Personalized Energy Management and Awareness Services for Energy Efficiency in Smart Buildings," *Sensors,* vol. 17, (9), pp. 2054, 2017. DOI: 10.3390/s17092054.

[6] J. Aguilar, A. Garces-Jimenez, M.D. R-Moreno, Rodrigo García, "A systematic literature review on the use of artificial intelligence in energy self-management in smart buildings", *Renewable and Sustainable Energy Reviews*, vol. 151, 2021, https://doi.org/10.1016/j.rser.2021.111530.

[7] L. Morales Escobar, J. Aguilar, A. Garcés-Jiménez, J. A. Gutierrez De Mesa and J. M. Gomez-Pulido, "Advanced Fuzzy-Logic-Based Context-Driven Control for HVAC Management Systems in Buildings," *IEEE Access*, vol. 8, pp. 16111-16126, 2020, doi: 10.1109/ACCESS.2020.2966545.

[8] G. Yoon, S. Park, S. Park, T Lee, S. Kim, H. Jang, S. Lee, S Park, "Prediction of machine learning base for efficient use of energy infrastructure in smart city," in *International Conference on Computing, Electronics & Communications Engineering (iCCECE)*, pp. 32-35, 2019.

[9] Q. Xiao, C. Li, Y. Tang, X. Chen, "Energy Efficiency Modeling for Configuration-Dependent Machining via Machine Learning: A Comparative Study," *Tase,* vol. 18, (2), pp. 717-730, 2021. DOI: 10.1109/TASE.2019.2961714.

[10] Z. Wu and W. Chu, "Sampling strategy analysis of machine learning models for energy consumption prediction," in *IEEE 9th International Conference on Smart Energy Grid Engineering*, pp. 77-81, 2021, doi: 10.1109/SEGE52446.2021.9534987.

[11] O. A. Olanrewaju, "Predicting industrial sector's energy consumption: Application of support vector machine," in *IEEE International Conference on Industrial Engineering and Engineering Management* pp. 1597-1600, 2019, doi: 10.1109/IEEM44572.2019.8978604.

[12] W. Chu, L. Spinella, D. Shirley, P. Ho, "Effects of wiring density and pillar structure on chip package interaction for advanced cu low-k chips," in *IEEE International Reliability Physics Symposium*, 2020, doi: 10.1109/IRPS45951.2020.9128333.

[13] D. N. Darlis, M. Abdul Latip, N. Zaini, H. Norhazman "Random forest approach for energy consumption behavior analysis," in *IEEE Symposium on Industrial Electronics & Applications* 2020, doi: 10.1109/ISIEA49364.2020.9188072.

[14] N. Zhang and D. Shetty, "An effective LS-SVM-based approach for surface roughness prediction in machined surfaces," *Neurocomputing,* vol. 198, pp. 35-39, 2016. DOI: 10.1016/j.neucom.2015.08.124.

[15] A. M. Abdulshahed, A. Longstaff, S. Fletcher, A. Potdar, "Thermal error modelling of a gantry-type 5-axis machine tool using a Grey Neural Network Model," *Journal of Manufacturing Systems,* vol. 41, pp. 130-142, 2016. DOI: 10.1016/j.jmsy.2016.08.006.

[16] D. Kong, Y. Chen and N. Li, "Gaussian process regression for tool wear prediction," *Mechanical Systems and Signal Processing,* vol. 104, pp. 556-574, 2018. DOI: 10.1016/j.ymssp.2017.11.021.

[17] D. A. Bashawyah and S. M. Qaisar, "Machine learning based short-term load forecasting for smart meter energy consumption data in london households," in *IEEE 12th International Conference on Electronics and Information Technologies*, pp. 99-102, 2021, doi: 10.1109/ELIT53502.2021.9501104.

[18] J. Aguilar, J., M. Cerrada, F., Hidrobo, F. (2007). "A Methodology to Specify Multiagent Systems". Lecture Notes in Computer Science, vol 4496, pp. 92–10, 2007. https://doi.org/10.1007/978-3-540-72830-6_10.

[19] J. Aguilar, C. Salazar, H. Velasco, J. Monsalve-Pulido, and E. Montoya. "Comparison and Evaluation of Different Methods for the Feature Extraction from Educational Contents" *Computation,* vol. 8, 2020.

[20] J. Aguilar, "Definition of an energy function for the random neural to solve optimization problems", Neural Networks, vol. 11, pp. 731-737, 1998, https://doi.org/10.1016/S0893-6080(98)00020-3.

[21] A. C. Müller and S. Guido, *Introduction to Machine Learning with Python.* (First edition ed.) 2016.

[22] A. M. Bagirov, N. Karmitsa and S. Taheri, *Partitional Clustering Via Nonsmooth Optimization.* (1st ed.), Springer, 2020.

[23] E. Camargo, J. Aguilar, Y. Quintero, F. Rivas, and D. Ardila. "An incremental learning approach to prediction models of SEIRD variables in the context of the COVID-19 pandemic". *Health Technol.* vol. 12, pp. 867–877 (2022).

[24] D. Pelleg and A. Moore, "X-means: Extending K-means with Efficient Estimation of the Number of Clusters," *Machine Learning,* 2002.

[25] D. Pham, S. Dimov and C. Nguyen, "Selection of K in K -means clustering," *Institution of Mechanical Engineers Part C-Journal of Mechanical Engineering Science,* vol. 219, pp. 103-119, 2005. DOI: 10.1243/095440605X8298.

[26] L. Morales and J. Aguilar, "An Automatic Merge Technique to Improve the Clustering Quality Performed by LAMDA," *IEEE Access,* vol. 8, pp. 162917-162944, 2020. DOI: 10.1109/ACCESS.2020.3021675.

[27] L. Morales, C. Ouedraogo, J., Aguilar, C. Chassot, S. Medjiah, K. Drira, "Experimental comparison of the diagnostic capabilities of classification and clustering algorithms for the QoS management in an autonomic IoT platform". *SOCA,* vol. 13, pp. 199–219, 2019.

[28] L. Morales Escobar, J. Aguilar, A. Garcés-Jiménez, J. A. Gutierrez De Mesa and J. M. Gomez-Pulido, "Advanced Fuzzy-Logic-Based Context-Driven Control for HVAC Management Systems in Buildings," *IEEE Access*, vol. 8, pp. 16111-16126, 2020, doi: 10.1109/ACCESS.2020.2966545.

[29] M. Mizumoto, "Pictorial representations of fuzzy connectives, Part I: Cases of t-norms, t-conorms and averaging operators," *Fuzzy Sets and Systems,* vol. 31, (2), pp. 217-242, 1989. DOI: 10.1016/0165-0114(89)90005-5.

[30] F. A. Ruiz *et al*, "A new criterion to validate and improve the classification process of LAMDA algorithm applied to diesel engines," *Engineering Applications of Artificial Intelligence,* vol. 60, pp. 117-127, 2017. DOI: 10.1016/j.engappai.2017.02.005.

[31] C. Bedoya, C. Uribe and C. Isaza, "Unsupervised feature selection based on fuzzy clustering for fault detection of the Tennessee Eastman process," in *Advances in Artificial Intelligence*: Springer, pp. 350-360., 2012

[32] M. Royapoor, M. Pazhoohesh, P. Davison, C. Patsios, S. Walker, "Building as a virtual power plant, magnitude and persistence of deferrable loads and human comfort implications," *Energy and Buildings,* vol. 213, pp. 109794, 2020. DOI: 10.1016/j.enbuild.2020.109794.

[33] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics,* vol. 20, pp. 53-65, 1987. DOI: 10.1016/0377-0427(87)90125-7.

[34] D. L. Davies and D. W. Bouldin, "A Cluster Separation Measure," Tpami, vol. PAMI-1, (2), pp. 224-227, 1979. DOI: 10.1109/TPAMI.1979.4766909.

[35] T. Papaioannou, G. Stamoulis, "Teaming and competition for demand-side management in office buildings," in *IEEE International Conference on Smart Grid Communications (SmartGridComm)*, pp. 332-337, 2017.

[36] "Power consumption data of a hotel building" https://ieee-dataport.org/documents/power-consumption-data-hotel-building

[37] L. Zhang, J. Wen, "Data for: A Systematic Feature Selection Procedure for Short-term Data-driven Building Energy Forecasting Model Development", Mendeley Data, https://data.mendeley.com/datasets/r532stprhv/1

[38] L. Zhang, J Wen, "A systematic feature selection procedure for short-term data-driven building energy forecasting model development", Energy and Buildings, vol. 183, pp. 428-442, 2019.

[39] M. Pipattanasomporn, G. Chitalia, J. Songsiri, J., Aswakul C., Pora W., Suwankawin S., Audomvongseree K. Hoonchareon N., "CU-BEMS, smart building electricity consumption and indoor environmental sensor datasets". Sci Data, vol. 7, 2020.

[40] Long-term energy consumption & outdoor air temperature for 11 commercial buildings, https://trynthink.github.io/buildingsdatasets/show.html?title_id=long-term-energy-consumption-outdoor-air-temperature-for-11-commercial-buildings