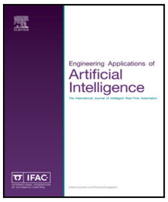
Engineering Applications of Artificial Intelligence 129 (2024) 107653

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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



Research paper

A novel clustering based method for characterizing household electricity consumption profiles

Francisco Rodríguez-Gómez, José del Campo-Ávila [*], Llanos Mora-López

Universidad de Málaga, Departamento de Lenguajes y Ciencias de la Computación, Campus de Teatinos, 29071 Málaga, Spain

ARTICLE INFO

Keywords:
Typical load profiles (TLP)
Big data clustering
Non-unique cluster numbers
Expert knowledge integration

ABSTRACT

A new methodology based on expert knowledge and data mining is proposed to obtain data-driven models that characterize household consumption profiles. These profiles are useful for electricity marketers to understand their customers' consumption. They could then adjust their electricity purchases in the market and provide recommendations to their customers to manage their consumption. The novelty of this research work is proposing a new procedure to determine an adequate number of clusters for a clustering task. Therefore, the proposed new methodology includes this novel procedure to build the models in two phases. In the first phase, clustering algorithms are used to group the data using different numbers of clusters. For the second phase, a new procedure (k-ISAC_TLP) is proposed and used to select the most appropriate number of clusters. This methodology allows the inclusion of domain information. In the case of household electricity consumption, where only groups with a significant number are relevant as long as the error is small, it allows the use of metrics like the mean absolute error and the number of observations (daily electricity consumption profiles). According to experts, the results achieved in two real datasets (from Spain and Ireland), with millions of observations support the methodology and reveal novel knowledge. In both cases, two and a half million observations have been analyzed and around twenty electricity consumption profiles have been detected. The methodology is easily extensible to problems of any domain where clustering algorithms are applicable. A software solution has been implemented and made freely available.

1. Introduction

Nowadays, in order to discover hidden knowledge to deal with complex problems involving large amounts of data and variables, data mining techniques have proven to be able to do so accurately and in a short time (Hastie et al., 2009). The significant increase in the amount of data and information available in recent years has led to the development of new data mining techniques to manage all this information and allow automatic extraction of valuable and relevant information.

Every data mining process usually consists of a first phase in which the problem to be treated and the necessary data are studied (Martinez-Plumed et al., 2021). Next, these data are processed and prepared, the models are generated, and later they are evaluated to check if the initial objectives have been achieved. The participation of experts in the entire data mining process is essential to obtain the expected results. Without the participation of experts, it would not be possible to discover if any model should be discarded or if it is relevant, nor if the data are the correct or if the problem is correctly understood.

Electricity consumption data is a real example where smart meters register vast amounts of data, and some different proposals transform them into useful information (Rajabi et al., 2020). Individuals can adapt and improve their lifestyle using the information that they themselves generate, but it is also possible to obtain the consumption profiles of domestic consumers from a global perspective. Discovering profiles, also called patterns, from electricity consumption data is a data mining task that will accelerate the transition towards future supply models, energy uses, sustainability and competitiveness in the medium and long term.

Having models that characterize the domestic hourly electricity consumption profiles provides advantages to the main players in this domain: (a) electricity marketers will be able to offer better tariffs and personalized services to their customers, as well as being able to better manage the electricity grid; (b) consumers could be advised to shift their consumption to times when electricity prices are lower (or null if it is produced by self-consumption facilities) or there is less demand; and (c) the environment would be positively impacted by

[*] Corresponding author.
E-mail addresses: francisco.rdg.gmz@uma.es (F. Rodríguez-Gómez), jcampo@uma.es (J. del Campo-Ávila), llanos@uma.es (L. Mora-López).

https://doi.org/10.1016/j.engappai.2023.107653
Received 8 February 2023; Received in revised form 19 August 2023; Accepted 29 November 2023
Available online 12 December 2023

List of abbreviations

CDI	Clustering Dispersion Indicator
$CKSC$	Combination of k-means and Spectral Clustering
CVI	Clustering Validity Indices
DB	Davies–Bouldin index
FCM	Fuzzy c-means
GMM	Gaussian Mixture Model
$ISAC$	Identifier of Stable Areas in Curves
$k-ISAC\_TLP$	k determination via ISAC method for TLP
MAE	Mean Absolute Error
MIA	Mean Index Adequacy
MSE	Mean Squared Error
PC	Profile Classes
SIL	Silhouette index
SOM	Self Organized Maps
TLP	Typical Load Profile
$WCSS$	Within-Cluster Sum of Squares

better management and use of the electricity grid, enabling the fight against climate change, and in improving urban sustainability.

1.1. Related work on finding domestic electricity consumption profiles

The use of machine learning methods to discover consumer profiles is a reality for some years now. Since the beginning of the century, there are academic works that propose the use of the Kohonen self-organizing map (SOM) to find customer clustering using a few tens or hundreds of consumers (Figueiredo et al., 2003; Verdu et al., 2004). However, despite the widespread use of SOM modeling, there are other algorithms that achieve better performance (Rajabi et al., 2020).

More recent proposals, like the one presented by Räsänen et al. (2010), continue using the same approach, but considering thousands of consumers instead of the hundreds used in previous works, which enriches the dataset and makes it possible to find a greater variety of profiles. The authors combine some clustering methods (such as k-means and hierarchical clustering) with SOM. But even with improvements in HW, the computational limitations require a reduction of the dataset. Thus, they process the raw data, reducing the length of the initial time-series and retaining only 5%. To decide the number of clusters to configure the k-means algorithm, they calculate the Davies–Bouldin index and the average within cluster variance.

In other work, several load Profile Classes (PC) are proposed to characterize the domestic customer electricity consumption (McLoughlin et al., 2015). This PCs are then used to make a classification of customers according to the sequence of consecutive PCs defining each customer for six months. In the first step, where profiles have to be discovered, unsupervised learning methods are used to characterize diurnal, intra-daily and seasonal patterns of electricity use. The data are taken from a dataset where 5000 Irish homes and businesses participated (Irish Social Science Data Archive – ISDDA – (Commission for Energy Regulation (CER), 2012)). This dataset, due to its availability, is used in numerous works. This is the case of Sun et al. (2020) where a new method (CKSC) that combines k-means and spectral clustering is proposed. In this work, the Silhouette index (SIL) is used to support the determination of the appropriate number of clusters, being one of the most recently used metrics in this field.

In the review by Cembranel et al. (2019), a summary of the methods used for clustering load electricity profiles is presented. They conclude that the k-means algorithm presents the best results, although they point out the problem of deciding the correct number of clusters. To avoid this issue, the authors suggest using automatic algorithms, such as G-means (Hamerly and Elkan, 2003) or X-means (Pelleg and Moore, 2000). These algorithms run the k-means algorithm in succession until the acceptance of a condition stops the search, but they also present limitations. For example, G-means is used in the domain of characterizing electricity consumption (Mets et al., 2016) but original data has to be approximated by a wavelet representation. The reduction of dimensionality is needed because G-means requires not so highly dimensional data. Regarding X-means, it is slow for large amounts of data.

Another proposal that also faces the same dimensionality problem is made by Yilmaz et al. (2019). They use only five features to define the shape of household electricity consumption, instead of using all values along the day.

A similar approach that describes the demand for every consumption with seven calculated features is proposed by Kaur and Gabrijelčič (2022).

In this area, the so-called "curse of dimensionality" problem appears, it is well known in the field of machine learning. This problem affects both the number of features that define an observation and the number of observations themselves.

Several clustering techniques (k-means, fuzzy C-means, hierarchical clustering and SOM) used in the context of electrical load patterns are compared by Rajabi et al. (2020). The clustering validity indices (CVI) used are Mean Square Error (MSE), Silhouette (SIL), Davies–Bouldin (DB), Mean Index Adequacy (MIA), Dunn, and the ratio of within-cluster sum of squares to between-cluster variation (WCSS). The main source of data is again the Irish dataset (ISDDA) where some filters and transformations are applied to reduce the dimensionality. After comparing the performance of clustering methods and cluster validity indexes (CVIs) they propose different number of clusters depending on the size and nature of the datasets built.

The growing use of smart meters is accelerating the emergence of really large datasets that record hourly energy consumption (millions of records). But the computational requirements of clustering algorithms limit the number of records and variables that can be processed using standard amounts of time and memory. One alternative is selecting a subset from the whole dataset. This is the approach proposed by Kwac et al. (2014) that starts with tens of millions of records and work with one hundred of thousands (by selecting households in the same location or by random sampling). The first stage in their methodology is the creation of a load shape dictionary for representative load shapes. They generate a dictionary with approximately 5K clusters for one subset with about 140K load patterns (taken from a specific zip code). Then the dictionary is reduced to 1K clusters that covers the 95% of all load shapes over all the areas and periods. They assure that the representative load shapes are consistent regardless of spatial and temporal locality, but they do not check the consistence of different dictionaries (clusters for load shapes) created from different subsets (from other zip codes).

Toussaint and Moodley (2020) emphasize that selecting a useful set of clusters to identify dominant electricity consumption profiles of households requires extensive experimentation and domain knowledge. Experts' knowledge is important in this process and the authors address this need by formalizing implicit expertise as external evaluation measures. Thus, the evaluation of results is measured by using internal metrics (like mean index adequacy, Davies–Bouldin index, the Silhouette index or a new Combined Index) and external metrics (like mean demand error or mean peak coincidence ratio). In addition to their proposal, they also include an analysis of the most commonly used clustering algorithms for this task in 25 academic works. K-means is both the most used algorithm and the one that ends up achieving the best results most frequently. This conclusion is also supported by their own results, because they verify that k-means is the best clustering algorithm between all experiments they conducted.

This algorithm, k-means, is the selected option in very recent works that has the discovery of relevant profiles as a crucial step in its methodology. Guo et al. (2022) and Rafiq et al. (2023) need to determine those profiles, based on hourly consumption, because they try to predict residential electricity consumption. It is an advanced task, but one that relies for its development on a good determination of household intraday electricity consumption profiles.

Losing variability is another issue present in this area. Aforementioned works by Kwac et al. (2014) or Yilmaz et al. (2019) show it. The latter incorporated to the study the average household electricity profiles (aggregated values) and detected 3 distinct clusters, instead of the 4 clusters detected when not using aggregated data. They conclude that averaging the data suppresses the diversity of the electricity use profiles within the individual household, while demonstrating that the raw daily profiles for a household differ significantly from its average profile.

The most relevant data on datasets, algorithms and results of the referenced research papers are summarized in Table 1. This summary highlights the progress made to date, as well as identifying several gaps that this work aims to fill:

- There are many configurations for carrying out the clustering task (algorithms, distances, CVIs). K-means appears to be the most used clustering algorithm in this context, maybe because its low complexity (Xu and Tian, 2015), and the one that achieves better results (Toussaint and Moodley, 2020). It is also an easily parallelizable algorithm. Based on CVIs, the Silhouette index, used in recent works in the area, can be used to assess the quality of clusters, but sometimes tends to choose a smaller number of clusters, which is not suitable for profile segmentation (Guo et al., 2022). Combining this index with some other criteria could improve the determination of the number of clusters.
- The dimensionality of the dataset, either by the number of instances or by the number of variables defining them, is one of the major limitations when running clustering algorithms. It is relatively easy to learn from thousand of daily load profiles, but the task becomes much more complicated when millions are involved. It is usual to sample the dataset to use approximately 100K profiles at most. Considering 24 variables, one per hour in a day, is also very common. Only recent works have used millions of observations to group similar profiles in the same cluster (Toussaint and Moodley, 2020; Guo et al., 2022). It is a time consuming task that can be accelerated by using parallel computing.
- Determining a single optimal number of clusters (k) is a task that is tackled from two different perspectives: considering multiple values for k and getting the best value a posteriori (checking CVIs measures) or doing a succession of executions until the acceptance of a condition stops the search. The sizes proposed in different research works are very heterogeneous: from a few clusters (less than 10) to many of them (even thousands) through some tens (one to three tens). One of the main factors for this variability may be the inclusion of experts' knowledge. This highlights the importance of developing methodologies that allow for the inclusion of experts' experience. What seems to be agreed is that, in general for this context, working with few clusters (less than ten) is insufficient, but working with too many (more than thirty) over-complicates the options being interpreted by the experts.
- It is important to use non-aggregated daily load profiles to avoid the loss of variability observed in domestic load consumption (Yilmaz et al., 2019). With the same intention of not conditioning the results to be achieved, there are several strategies that could be eliminated. For example, it is common to segment the data prior to the clustering phase (weekday and weekend (Rajabi et al., 2020), cooling-included and cooling-excluded (Rafiq et al., 2023)), but biases can be introduced. Normalization is another common phase in the preparation of the data (Toussaint and Moodley, 2020; Kaur and Gabrijelčič, 2022), which, among other objectives, allows to agglutinate patterns that share the same shape and only differ in scale, but hinders a correct retrieval of the original patterns.

Taking into account the progress in the characterization of household electricity consumption made to date, the limitations detected and the opportunities, some research questions arise:

- Can the Silhouette index be complemented with some other criteria to improve the determination of the number of clusters?
- Moreover, can an automatic method be proposed to determine appropriate values for the number of profiles to be searched for?
- Is using a parallelizable version of k-means an advantageous alternative for working with large volumes of data?
- Is there any advantage in simplifying the pre-processing step before clustering the dataset?

1.2. Contributions and organization of the paper

In this paper, we present the advances achieved by combining the experts' knowledge inside a data mining process to characterize the domestic hourly electricity consumption profiles. Some contributions can be perceived from different perspectives, but the most relevant are the following:

- A new methodology that automatically allows characterizing the domestic hourly electricity consumption profiles through data-driven models. Even with millions of daily load profiles, viable alternatives, like big data algorithms, take advantage of computational resources. An implementation of this methodology is made available as open-source. This methodology simplifies the pre-processing data steps.
- A new procedure to determine non-unique cluster numbers based on the stability of different metrics recorded during the implementation of the methodology. Those candidates of the number of clusters can then be used to select the most appropriate one. This proposal can incorporate experts' knowledge depending on the domain problem.
- The domestic hourly electricity consumptions for two different locations are characterized by using the proposed methodology. They have been identified using clustering methods and considering knowledge provided by experts. Experimental results and expert advice suggest that this methodology is a valuable tool. One of the datasets is public and available on request (Commission for Energy Regulation (CER), 2012).

These contributions have direct application in the real world. For example, the proposed procedure for determining the appropriate number of clusters, based on the stability of different metrics, can be applied to any real-world domain where profile characterization is useful, such as modeling electricity consumption profiles, modeling business study dimensions, or modeling environmental phenomena for climate change action. The profiles detected in the two localities can be used by local electricity companies to give consumption recommendations, shift consumption to cheaper hours or propose groups of consumers for possible energy communities.

The rest of the paper is organized as follows. Background knowledge used in this work is described in Section 2. It includes a description of the clustering techniques used and the metrics and statistical validation proposed for evaluating the models. The proposed methodology is explained in Section 3. The experimental design is described in Section 4, including the description of the dataset and some details about the implementation. Section 5 presents the intermediate results that guided the methodology to a final result in real cases context. This section also includes the evaluation done by experts in the domain. Finally, Section 6 summarizes the main conclusions of the work.

Table 1
Summary of datasets, algorithms and results shown in the referenced research papers. References are presented in ascending chronological order.

	Country	Observations	Algorithms	Cluster metric	Best k	Available dataset
Figueiredo et al. (2003)	Portugal	165	SOM k-means	CDI	9	–
Verdu et al. (2004)	Spain	327	SOM	–	5	–
Räsänen et al. (2010)	Finland	3989	SOM + k-means SOM + hierarchical	DB	19	–
Kwac et al. (2014)	USA	Sample 200K Consumers	Adaptive k-means + + hierarchical	MSE	1000 (16)	–
McLoughlin et al. (2015)	Ireland	3941	SOM k-means k-medoids	DB	10	✓
Yilmaz et al. (2019)	Switzerland	656	k-means	SIL	4	–
Sun et al. (2020)	Ireland	Sample 4181 Consumers	CKSC clust.	SIL	6+6	✓
Rajabi et al. (2020)	Ireland	Reduce Data from 4141 Consumers	SOM k-means Hierarchical FCM clust.	DB SIL MIA ...	16 + 16	✓
Toussaint and Moodley (2020)	South Africa	3.3MM	SOM k-means SOM + k-means	DB + + SIL + + MIA	59	✓
Kaur and Gabrijelčič (2022)	Slovenia	4963	GMM clust.	mix	15	–
Guo et al. (2022)	Ireland	1.9MM	k-means	SIL	6	✓
Rafiq et al. (2023)	Dubai	134K	k-means	WCSS	4+4	–
k-ISAC_TLP	Spain	2.4MM	k-means	SIL	19	–
(this proposal)	Ireland	2.5MM	Bisecting k-means	MAE	21	✓

2. Background and preliminaries

Clustering is an unsupervised machine learning technique that is used to separate and group the observations of a dataset into different subsets. There are different applications, such as community detection (Berahmand et al., 2018) or customer segmentation (Figueiredo et al., 2003). The observations in each subset should be as similar as possible to each other in the subset, subject to a distance function. Additionally, it is desirable that the distance between different subsets is high.

In the context of household electricity consumption, there are no predefined profiles and it is important to identify how many typologies of consumption exist. Therefore, unsupervised learning is needed and clustering becomes an appropriate tool to find those profiles.

In this section we describe the most relevant aspects to be considered for conducting a clustering task: distance between observations, algorithms, indicators to evaluate the suitability of the model and determination of appropriate number of clusters. In Section 3.2 we complement with new proposals those described in this section.

2.1. Distance functions

A distance function is needed to measure the similarity between observations described by quantitative attributes. Note that the clustering process consists of grouping similar observations, so a form of measurement is needed.

One of the most common functions is the Euclidean distance. In case of considering m variables (or attributes) to define an observation it can be formulated (in m-dimensional Euclidean space) as:

$$euclidean\_dist(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_m - y_m)^2} \qquad (1)$$

where x and y are two points in a m-dimensional space ($\mathbb{R}^m$).

When calculating the error of fitting the whole dataset by a model, the sum of the squared Euclidean (SSE) distances between each observation and the centroid of the corresponding cluster is used to get the Mean Distortion (MD) in the whole model (Shi et al., 2021):

$$SSE = \sum_{i=1}^{k} \sum_{o_j \in C_i} euclidean\_dist(o_j, centroid_i)^2 \quad ; \quad MD = \frac{SSE}{n} \qquad (2)$$

where k is the number of clusters, n is the number of observations, o_j is an observation ($j \in [1, n]$) assigned to the cluster i (C_i where $i \in [1, k]$) and $centroid_i$ is the centroid of the cluster i.

Since each observation in this domain is the hourly consumption of one day ($m = 24$), another types of distances could take into account the form of the curves like Dynamic Time Warping (DTW) distance (Sakoe and Chiba, 1978) or the kernel distance (Cuturi, 2011).

2.2. Clustering methods

There is a wide range of clustering algorithms (Xu and Tian, 2015) that can be classified according to how they perform the clustering: partitionally, hierarchically, based on density, based on graph partition. Within all this variety of algorithms some of them are better prepared to work with massive datasets in different domains. Attending to time complexity, only algorithms with $O(n)$ or at most $O(n \cdot log(n))$ could be used:

- *Based on partition.* K-means (MacQueen, 1967) starts selecting k clusters each of which consists of a single random point (centroid). Thereafter k-means adds each observation in the dataset (n is the number of observations) to the cluster whose centroid is nearest. An update process continues changing the centroids until some criteria for convergence is met (t is the number of iterations). This procedure is of low time complexity ($O(knt)$), so that it is feasible to process very large samples.

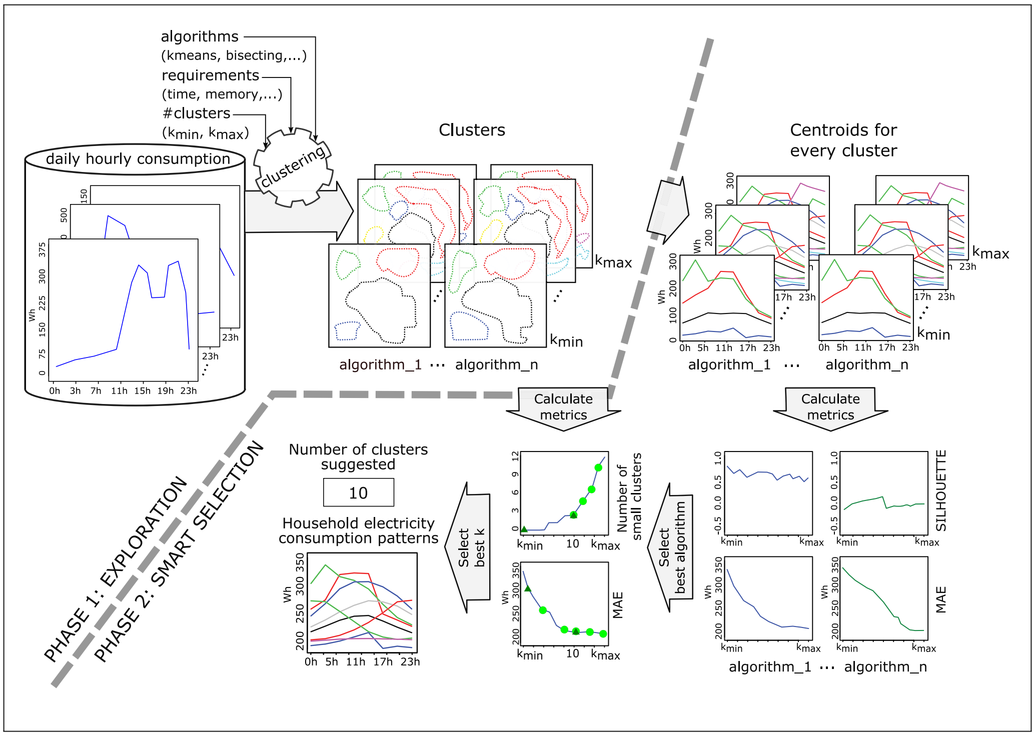


Fig. 1. Methodology proposed to detect household electricity consumption profiles from daily hourly consumption data. It is divided into two phases, the first one to explore different alternatives and the second one to propose one of them as a suitable model.

- *Based on hierarchy.* Bisecting k-means (Steinbach et al., 2000) is a divisive hierarchical clustering algorithm that uses k-means to refine successive divisions. Instead of partitioning the dataset into k clusters from the beginning, bisecting k-means splits one cluster into two sub-clusters at each bisecting step (by using k-means) until k clusters are obtained. Its time complexity is, like k-means, linear in the number of observations (n). If the number of clusters is large, bisecting k-means could be even more efficient than the classical k-means algorithm. Another approach with low time complexity ($O(n)$) is BIRCH (Zhang et al., 1996). It creates and dynamically updates a Clustering Feature Tree (CF tree), of which one node stands for a subcluster. That information is used to conduct an agglomerative hierarchical clustering algorithm that create the final clusters in the model.
- *Based on density.* DBSCAN (Ester et al., 1996) is the most well known clustering algorithm in this category. The basic idea is grouping together those observations of the data space that are in a region with high density. OPTICS (Ankerst et al., 1999) is an evolution of DBSCAN that incorporate two parameters (the radius of the neighborhood and the minimum number of points in a neighborhood) to improve the induction of the model. Both algorithms show slightly higher complexity than linear: $O(n \cdot log(n))$.
- *Based on graph theory.* Power Iteration Clustering (PIC) (Lin and Cohen, 2010) is a clustering technique based on graph theory. It requires a row-normalized affinity matrix and it can find the clustering partition using linear complexity in the number of observations ($O(n)$). Although its complexity is low, it assumes the existence of an affinity matrix. Previously to execute PIC algorithm, or any other based in the idea of partitioning the graph (like CLICK (Sharan et al., 2003)), an undirected weighted graph must be constructed (Dhanapal and Perumal, 2016). Every observation in the dataset is a vertex and the similarity value between any two observations is the weight of the edge connecting the two vertices. Generating this matrix triggers the time complexity ($O(n^2)$).

2.3. Evaluation metrics

To evaluate the quality of the clustering models, there are many indices, but only internal evaluation is possible when there is none information about the real groups. Internal indicators usually evaluate how similar the observations of the same cluster are (cohesion), and how different they are with respect to the observations in other clusters (separation). There is a wide diversity, but the most common metrics used when clustering electricity consumption profiles are the following:

- *Silhouette index* (Rousseeuw, 1987) is a metric that combines (a) the average dissimilarity of one observation to all other objects of its own cluster (cohesion); and (b) the average dissimilarity of one observation to all other objects of its most closest cluster (separation). This combination results in the calculation of the index for every observation $s(o_j)$. When this index is aggregated for every observations in a cluster and then aggregated for all the clusters, the overall average Silhouette width is obtained. This overall value ranges between [−1,1], where higher values indicates an appropriate modeling of the clusters.
- *Davies–Bouldin index* (Davies and Bouldin, 1979) measures the similarity of clusters which are assumed to have a data density which is a decreasing function of distance from a vector characteristic of the cluster (centroid). It calculates, for every pair of clusters, the ratio between (a) their intra-cluster dispersion (or cohesion); and (b) the distance between their centroids (separation). Then averages the maximum ratios for every cluster. Davies–Bouldin index ranges between [0,1], where lower values mean better clustering models.

2.4. Determining the number of clusters

Finding how all data should be correctly assigned to different clusters is a fundamental problem even when the optimal number of clusters is known. But this problem is aggravated when that number is unknown (Ezugwu et al., 2021; Rostami et al., 2023). There exist different alternatives to look for the appropriate numbers of clusters. The two main approaches are: (a) creating clustering models with different values of k in a predefined range ($k \in [k_{min}, k_{max}]$) and subsequently determining the best (unique or non-unique) values; and (b) building a succession of clustering models, with increasing k, until the acceptance of a condition stops the search.

There is a great diversity of methods and algorithms following the first approach. One of the simplest is based on the aforementioned Silhouette index values, since one way to choose k is to select the value that results in a higher value of the overall average Silhouette width. Another idea is known as the elbow method, that considers the best number of clusters to be that in which there is no longer a significant change between the value of the index for that number of clusters and the value of the index for the next proposed number of clusters. This identification is manually done by visualizing a curve, which can cause situations where experts cannot clearly identify the elbow point (smooth curves). To overcome this difficulty, Shi et al. (2021) propose the selection of the value of k where the minimum angle between consecutive values is observed. The combination of the elbow method and Silhouette metric is one alternative (Raj and Vidyaathulasiraman, 2021) that attempts to minimize odd situations. The idea implemented in NbClust package (Charrad et al., 2014) goes beyond and calculates several metrics (even tens of metrics) and selects the value of k voted by the majority. They argue that it is difficult to reach a unanimous decision on the optimal number of clusters.

With regard to the second approach, there are also multiple alternatives. In Section 1.1 it is described that G-means and X-means have been used in the context of domestic electricity consumption profiles. Without restricting the scope to this domain, Ezugwu et al. (2021) make a review where some other dynamic clustering algorithms are compiled. But, independently of the domain, these authors conclude that there is a big issue with large-scale datasets especially when handling real-world clustering problems.

3. Methodology and new proposals

This section presents the proposed methodology to characterize the domestic hourly electricity consumption profiles. Moreover, a new proposal to determine the optimal number of these profiles is also described.

3.1. Methodology

The process of characterizing consumption profiles is divided into two phases. Fig. 1 represents these phases and how they combine to induce the model that summarizes the knowledge in the available data.

- The first phase carries out an exploratory search guided by the experts in the domain of electricity consumption and supported by data scientists. Experts define the range for the number of clusters ($k \in [k_{min}, k_{max}]$). They know that there exist a minimum number of different profiles and estimate that considering a large number of profiles would not provide useful information (usually irrelevant and small groups are created). Data scientists conduct different experiments with available algorithms and methods in the range defined by the experts, assuming realistic conditions (limiting available time and memory).

  Depending on the size of the dataset, the variety of algorithms that can be used changes. If the dataset is large, options with time complexity beyond linear will not properly work. Even those with the lowest time complexities can have limitations if traditional implementations are used. In the case of highly demanding scenarios, big data clustering methods can be used (Dafir et al., 2021), although there are far fewer options (Saeed et al., 2020). Due to the lack of clustering algorithms that are able to cope well with the high computational cost (Ezugwu et al., 2021), new algorithms will be developed and they can be included in the repository of methods available in the proposed methodology.

  In this first phase, the Euclidean distance (Eq. (1)) is used to assign the observations to the clusters. The assignment is done by seeking to minimize this distance, because this minimizes the mean distortion (Eq. (2)).
- In the second phase, the model that best characterizes the hourly electricity consumption profiles is selected. Several stages go extracting the relevant information and propose the model with most appropriate k clusters.

  - Once the different clustering models have been generated, the centroids for every cluster are calculated. This step allows the calculation of the distance between observations and centroids with a distance function different from the usual Euclidean distance.

    Since each observation in this domain is the hourly consumption of one day ($m = 24$), the most appropriate distances could take into account the form of the curves. Although alternatives exist in the time series domain based on the shape of the curve (like DTW), experts in household electricity consumption have suggested another measurement: the Absolute Error (AE). Comparing two shapes of equal length (m variables) is quadratic for DTW ($O(m^2)$), but calculating the Absolute Error (AE) takes linear time ($O(m)$).

    $$AE(x,y) = |x_1 - y_1| + |x_2 - y_2| + \cdots + |x_m - y_m| \tag{3}$$

    where x and y are two points in a m-dimensional space ($\mathbb{R}^m$). When calculating the error of the whole dataset using absolute error leads naturally to the Mean Absolute Error (MAE) measure:

    $$MAE = \frac{\sum_{i=1}^{k} \sum_{o_j \in C_i} AE(o_j, centroid_i)}{n \times m} \tag{4}$$

    where k is the number of clusters, n is the number of observations, m is the number of variables, o_j is an observation ($j \in [1,n]$) assigned to the cluster i (C_i where $i \in [1,k]$) and $centroid_i$ is the centroid of the cluster i.

    Including the experts' knowledge, we follow the recommendation to align cluster evaluation measures with the specific objective of the application because it is essential to generate clusters that are helpful (Aggarwal, 2015).

    For experts, it is important to include information about the observed error in the model to validate its usefulness. Therefore, the calculation of MAE (Eq. (4)), which uses the definition of AE (Eq. (3)), provides relevant information on this characteristic.

    In addition to MAE values calculated for every model with different k values, Silhouette indexes are also calculated. MAE and Silhouette metrics are plotted separately depending on the algorithm and presented to the experts.
  - Taking into consideration the information calculated before (MAE and Silhouette), the experts, with the advice of data scientists, select the algorithm that best groups the observations. The main criterion is minimizing the MAE values (minimizing the error), but in case of having similar behavior, the Silhouette indexes can help in the decision.
  - Finally, determining the most appropriate number of clusters is crucial. Besides a unique proposal for such a number (most methods find only one value for k), it is possible to suggest additional proposals, resulting in non-unique cluster numbers (Borjigin and Guo, 2013). In this methodology, a new procedure to determine non-unique cluster numbers and identify the most suitable one is proposed; it is called k-ISAC-TLP and it is based on an also new method called ISAC. Although both of them are described in the following

Section 3.2, the general idea is given now to complete the methodology description. ISAC method uses the shape of curves and looks for regions of stability and absence of improvement. In this final stage, two curves are used: the MAE for different k values (calculated in the previous stage) and a new metric that counts the number of small clusters (and potentially useless) present for different k values. When combining the information obtained from both curves, a set of candidate values is given, and an appropriate value can be selected from this set.

3.2. New procedure for determining an appropriate number of clusters

As discussed in Section 2.4, there are different methods to assess the identification of the most appropriate (or optimal) number of clusters to consider while performing a clustering task. One of the most commonly used is the elbow method (a manual and visual approach), which looks for points on a curve where there is no longer a significant improvement (the "elbow" of a curve) (Rafiq et al., 2023). But this method suffers from some disadvantages, such as working with smooth or noisy curves (where it is difficult to identify the "correct" elbow) or the uncertainty of detecting a premature point (when there is some room for improvement).

This Subsection details a procedure called k-ISAC_TLP, designed to automate detecting the most appropriate number of clusters, avoiding visual interpretation and preventing early detection (see 3.2.2). This procedure is based on a new method, called ISAC (Identifier of Stable Areas in Curves), which attempts to identify areas in a curve with low variability in the values while negligible improvement in the metric measured by such values (see 3.2.1). These areas usually appear shortly after the elbow in a curve, so this search meets the needs of experts who wish to avoid premature search stoppage and prefer more conservative detections.

3.2.1. ISAC method (Identifier of Stable Areas in Curves)

The main idea of this method is to construct consecutive triangles along the path defined by a curve. From these triangles it is possible to obtain relevant information: the areas of the triangles and the slopes defined by the farthest vertices of the triangles.

Intuitively, the area of a triangle can show some information about the alignment of the points that define its vertices: if these points are aligned, the area tends to be zero. Therefore, finding triangles with small areas will suggest the existence of stable regions in the curve. A similar idea has been previously used for identifying stabilization in the learning curve (Castillo and Gama, 2006; del Campo-Ávila et al., 2008).

In the same intuitive manner, the slope between the two farthest vertices in the triangle can indicate the velocity of change in the curve: a slope close to zero suggests small progression. This type of progression has a different interpretation depending on the metric. If the curve describes a metric that relates the complexity of a model and a metric that should be minimized (like error), a region with a small slope indicates that little improvement is expected even with more complex models (possible stop point). Reversely, if the curve describes a metric that worsens after an initial plateau (like the number of irrelevant rules in a model), a slope with a significant value indicates the starting of a degradation (possible stop point).

In Fig. 2 different curves are presented in which the casuistry mentioned above can be appreciated: triangles with different values for the area and different slopes between the farthest vertices of the triangles.

The goal of the ISAC method is to find the points where: (a) a series of consecutive triangles have an area less than or equal to a maximum area ($areaThreshold$) , and (b) a slope greater than a minimum slope ($slopeThreshold$). The definitions for calculating these areas and slopes are given below.

Let p, q and r be three points that define a triangle and p_x, p_y, q_x, q_y, r_x and r_y their coordinates for X-axis and Y-axis. The function that calculates the area for that triangle is:

$$area(p,q,r) = \left| \frac{1}{2} \cdot ((q_x \cdot p_y - p_x \cdot q_y) + (r_x \cdot q_y - q_x \cdot r_y) + (p_x \cdot r_y - r_x \cdot p_y)) \right| \tag{5}$$

Let p and r be the two farthest vertices in a triangle. The function that calculates the slope (in degrees) is:

$$slope(p,r) = \arctan\left(\frac{r_y - p_y}{r_x - p_x}\right) \cdot 180^\circ/\pi \tag{6}$$

where $arctan$ is the function that calculates the arc whose tangent has the value of the passed parameter.

The area calculation (Eq. (5)) is used in Algorithms 1 and 3 when executing the $calculateArea(p,q,r)$ method. Symmetrically, the slope calculation (Eq. (6)) is used in Algorithms 1 and 4 when executing the $calculateSlope(p,r)$ method.

Besides the area and slope thresholds, some other parameters needed must be configured, like the size of the triangles or the number of consecutive stable triangles to avoid spurious results. A detailed implementation of ISAC method is given in Algorithm 1.

The size of triangles is determined by separating the vertices considering concrete values in the X-axis. Given a vertex in the X-axis (most left vertex in the triangle) and adding a distance twice (from the first to the second vertex and from the second to the third vertex), the triangle's vertices are defined. The distance concerned is called $distanceBetweenTrianglePoints$ in Algorithm 1. The other parameter, the number of consecutive stable triangles (called $consecutStability$ in the algorithm), is used to avoid spurious detections of stability in a region. The three vertices of a triangle may be well aligned (with a very small area in the triangle), but some instability can remain in points that do not define the triangle, but are between the vertices that define the triangle (only when $distanceBetweenTrianglePoints \geq 2$) . Identifying stability is more confident when multiple and consecutive small triangles are considered.

In summary, for the application of the ISAC method it is necessary to give the input data, that is, the curve to be analyzed (defined by x_values and $measure\_values$), and to set some parameters. These parameters define the size of the triangles ($distanceBetweenTrianglePoints$), the persistence of the condition that is fulfilled ($consecutStability$) and the thresholds for area and slope ($areaThrehold$ and $slopeThresold$).

3.2.2. k-ISAC_TLP procedure (k determination via ISAC method for typical load profiles)

The ISAC method is a multipurpose method that can be used for any curve. In this Subsection, it is customized for use in the field of domestic electricity consumption, where it is crucial to discover typical load profiles (TLP). The stabilization regions detected by ISAC can evaluate the determination of the most appropriate number of profiles (k) discovered by a clustering algorithm. Therefore, this procedure using ISAC has been named k-ISAC_TLP.

According to experts' knowledge there are two specific curves that may be significant in selecting useful consumption profiles. They are specifically defined as: (1) the Mean Absolute Error (MAE) of the clustering models for each value of k, and (2) the number of irrelevant clusters defined as those whose size is less than 1% of the whole dataset for each k value. In these curves the k values are arranged on the X-axis of the curve, and the measured values (MAE or number irrelevant clusters) are placed on the Y-axis.

k-ISAC_TLP procedure uses ISAC method to detect the most relevant points in each of the above curves. Non-unique cluster numbers are obtained for both curves resulting in values candidates to be useful. The combination of these non-unique cluster numbers can result in an outstanding proposal for the most appropriate k value. The following criteria is used:

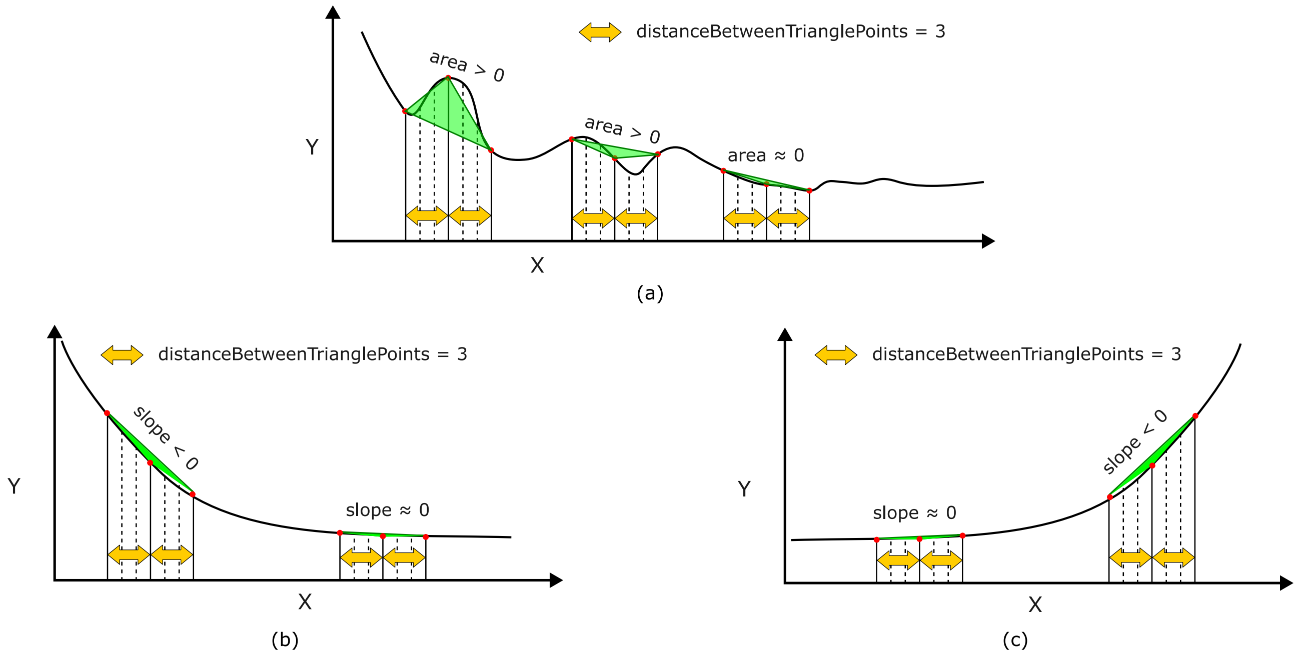

distanceBetweenTrianglePoints = 3
area > 0
area > 0
area ≈ 0
Y
X
(a)
distanceBetweenTrianglePoints = 3
slope < 0
slope ≈ 0
Y
X
(b)
distanceBetweenTrianglePoints = 3
slope < 0
slope ≈ 0
Y
X
(c)

Fig. 2. Visual description of calculated areas and slopes for triangles in a curve. (a) Curve with different sizes of areas for the triangles depending on the alignment of their vertices, (b) Curve that starts with negative slopes between the two farthest vertices in triangles and evolves until it approaches zero, (c) Curve that starts with slopes close to zero between the two farthest vertices in triangles and evolves to increase.

1st: minimum common number to both curves, if none
2nd: minimum number detected in MAE curve, if none
3rd: minimum number detected in counting small clusters curve.

The parameter configuration is the last issue before executing the ISAC method to complete the k-ISAC_TLP procedure. Considering the same configuration determined in Castillo and Gama (2006) and del Campo-Ávila et al. (2008) $distanceBetweenTrianglePoints$ is set to 3. By setting $distanceBetweenTrianglePoints = 3$, the ISAC method uses triangles that are neither too large nor too small. By considering $consecutStability = 3$ we avoid early and inconsistent detections, while reducing the overly stringent stability requirements that would require perfect scenarios. Regarding the area and slope thresholds, in the context of household electricity consumption, two different configurations have been selected depending on the curve:

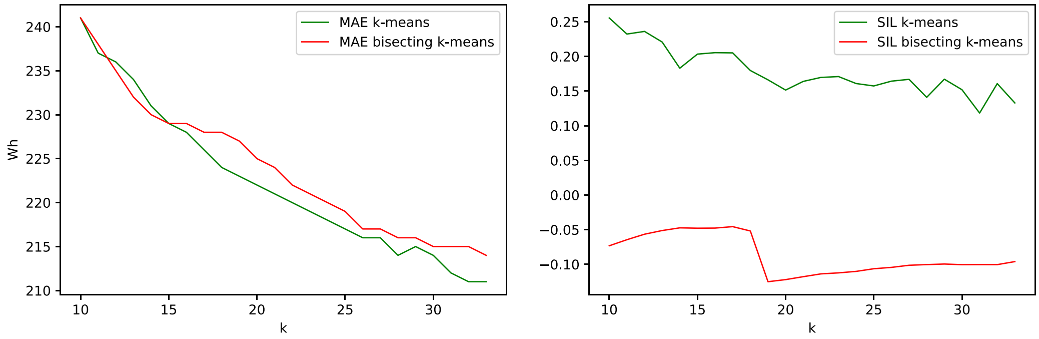
- *MAE curve*: the range of MAE values depends on the dataset and the error calculated when clustering it, so setting constant thresholds would not allow the procedure to be adaptive. The area and slope thresholds proposal should correspond with the whole curve. An estimation of the global behavior is extracted from a hypothetical triangle that starts at the beginning of the curve and finishes at the end (with an intermediate point in the middle of the curve). Therefore thresholds are defined as follows:
  - The area threshold ($areaThreshold$) is defined as the area of a triangle proportional to the hypothetical one, but considering the smaller size of triangles that will be used in the ISAC method.
  - The slope threshold ($slopeThreshold$) is defined as the slope between the first and last points in the curve.
- *Curve for the number of irrelevant clusters*: the range of this curve is known. It usually starts with a few irrelevant clusters (0 or close to 0) when considering the smallest k values, and, at some point, it usually starts increasing. The reason for that increment is the occurrence of small clusters that takes a few observations that could be integrated into another group. Still, those observations are separated unnecessarily from a bigger cluster because one (irrelevant) cluster is available and must be used.
Given that the behavior of this curve has been studied and given that the size of triangles used in the ISAC method has been previously defined, constant values have been assigned to thresholds as follows:
  - The area threshold ($areaThreshold$) is defined as 1.5, the next size value reachable after 0 (area values increase by 1.5 units as triangles get larger).
  - The slope threshold ($slopeThreshold$) is defined as 22.5°, that is half the 45° angle. When a model cannot be substantially improved by adding more clusters (increasing k), every new cluster would be irrelevant, and the number of irrelevant clusters will grow at the same rate that k value. That growth rate tends to have a 45° angle.

In summary, for the application of the k-ISAC_TLP procedure, detailed in Algorithm 2, it is only necessary to give the input data, that is, two curves to be analyzed (MAE_values and $irrelCluster\_values$). It is not necessary to set any parameters, as it is the procedure itself that defines the parameters to call the ISAC method. These parameters are set or calculated as follows: (a) $distanceBetweenTrianglePoints = 3$, (b) $consecutStability = 3$, (c) $areaThreshold$ is calculated for the MAE curve as a proportion of the triangle defined by the complete curve (see Algorithm 3), (d) $areaThreshold = 1.5$ for the number of irrelevant clusters curve, (e) $slopeThreshold$ is calculated for the MAE curve as the slope between the first and last points in the complete curve (see Algorithm 4), and (f) $slopeThreshold = 22.5$ for the number of irrelevant clusters curve. This configuration has been proposed for this domain after a close work with experts in the field.

4. Experimental design

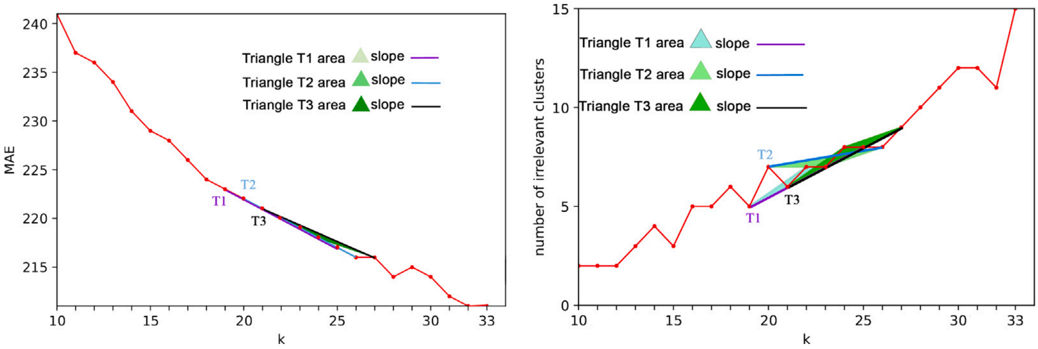
4.1. Datasets

Nowadays, household electricity consumption datasets are collected in private companies, but some are publicly available thanks to national



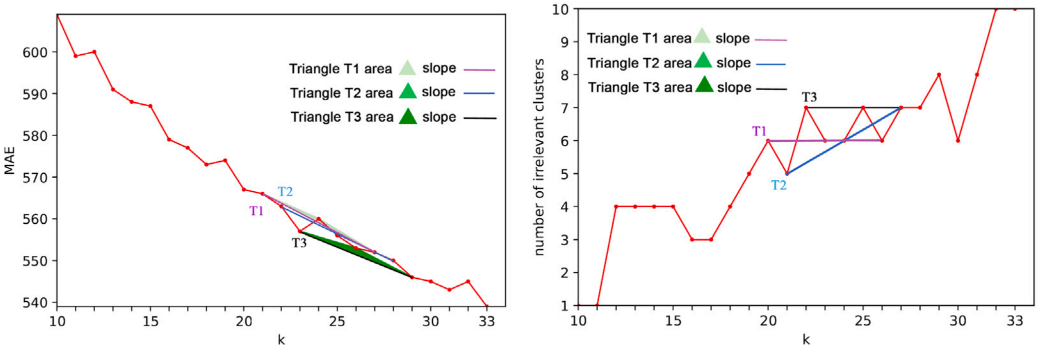
(a) MAE (Mean Absolute Error) (b) Silhouette index

Fig. 3. Comparison of the curves with the MAE and Silhouette index for the models generated with the k-means and bisecting k-means algorithms for the values of the range $k \in [10, 33]$ previously established by the experts. Lower MAE values are better. Higher Silhouette indices are better.
Data source: Spanish dataset presented in Section 4.1.



(a) Curve for MAE values (b) Curve for the number of irrelevant clusters

Fig. 4. Visual description of the automatic detection of k value proposed by the k-ISAC_TLP procedure for the model induced for the Spanish dataset. Best k value is 19, and the lowest k where both invocations to the ISAC method match. Starting from the value $k = 19$, for both curves, the triangles generated with this value of k and the two following ones have areas and slopes that satisfy the stopping conditions of the algorithm.



(a) Curve for MAE values (b) Curve for the number of irrelevant clusters

Fig. 5. Visual description of the automatic detection of k value proposed by the k-ISAC_TLP procedure for the Irish dataset. Best k value is 21, the lowest k detected by the ISAC method for the MAE curve. In this case there is no match between the curves, so priority is given to the MAE curve as defined in the methodology.

```
Algorithm 1: ISAC method
Input: // DATA: List with totalObs values for independent
       // variable x
       $x\_values[1, totalObs]$
       // DATA: List with measured values for totalObs
          measures
       $measure\_values[1, totalObs]$
       // number of values in X-axis between points of each
          triangle
       $distanceBetweenTrianglePoints$
       // number of consecutive triangles that have stable
          conditions
       $consecutStability$
       // minimum area value to consider that points in the
          triangle are aligned
       $areaThreshold$
       // maximum slope value to consider small progression
       $slopeThreshold$
Output: // x values where the curve satisfies area and
        // slope criteria
        $stable\_x\_values[]$

// Variables to store areas and slopes for every triangle built
   on the curve
$numberOfTriangles \leftarrow totalObs - (distanceBetweenTrianglePoints \cdot 2);$
$areaTriangles[1, numberOfTriangles] \leftarrow initialize;$
$slopeTriangles[1, numberOfTriangles] \leftarrow initialize;$
// for every triangle built in the curve
for $i \leftarrow 1$ to $numberOfTriangles$ do
    // Update vertices for current triangle
    $p_x \leftarrow x\_values[i]$ ;
    $p_y \leftarrow measure\_values[i];$
    $q_x \leftarrow x\_values[i + distanceBetweenTrianglePoints]$ ;
    $q_y \leftarrow measure\_values[i + distanceBetweenTrianglePoints];$
    $r_x \leftarrow x\_values[i + (distanceBetweenTrianglePoints \cdot 2)]$ ;
    $r_y \leftarrow measure\_values[i + (distanceBetweenTrianglePoints \cdot 2)];$
    // Calculate and store area and slope for current triangle
    $areaTriangles[i] = calculateArea(p, q, r);$
    $slopeTriangles[i] = calculateSlope(p, r);$
// for every triangle except last $consecutStability$ triangles
for $i \leftarrow 1$ to $(numberOfTriangles - consecutStability)$ do
    // Check for consecutive stable areas criteria
    $isStable \leftarrow true$ ; $j \leftarrow 0$ ;
    while $(isStable$ & $(j < consecutStability))$ do
        $isStable \leftarrow isStable$ & $(areaTriangles[i + j] \leq areaThreshold)$
                  & $(slopeTriangles[i + j] \geq slopeThreshold);$
        $j \leftarrow j + 1$ ;
    // Check for minimum slope criteria
    if $(isStable$ & $(slopeTriangles[i] \geq slopeThreshold))$ then
        $stable\_x\_values.add(x\_values[i])$ ;
```

organisms that support open data (Commission for Energy Regulation (CER), 2012; Toussaint, 2019). To test the validity of the methodology proposed in Section 3, two experiments have been carried out: the first is a private dataset of electricity consumption in southeastern Spain, and the second one is a public dataset from Ireland (Commission for Energy Regulation (CER), 2012). Selecting two datasets we can easily prove the methodology in two different regions with two different climates (a relevant aspect as experts refer). Climate in southeastern Spain is characterized by Mediterranean climate (hot dry summers and humid, cool winters) and its average annual temperature is 18°. Ireland has a temperate oceanic climate (warm summers, without dry season and cool winters) and its average annual temperature is 10°.

Both datasets register hourly data of household electricity consumption (Wh), so each observation has 24 values corresponding to the 24 hourly consumption of a consumer for a day. For the data preparation process the following filters suggested by experts have been applied:

```
Algorithm 2: k-ISAC_TLP procedure
Input: // DATA: values for $k$ to be considered in the range
       // defined by the experts
       $k\_values[k_{min}, k_{max}] = k_{min}, k_{min} + 1, ..., k_{max}$
       // DATA: List with MAE values measured for different $k$
          values
       // For simplicity, assume that indexes are named
          $k_{min}, k_{min} + 1, ..., k_{max}$
       $MAE\_values[k_{min}, k_{max}]$
       // DATA: List with the number of irrelevant clusters
          modeled for
       // different $k$ values
       // For simplicity, assume that indexes are named
          $k_{min}, k_{min} + 1, ..., k_{max}$
       $irrelCluster\_values[k_{min}, k_{max}]$
Output: // values for $k$ that meet the criteria defined
        // by experts
        $best\_k\_values[]$

// Configure parameters before calling ISAC method
$distanceBetweenTrianglePoints \leftarrow 3;$
$consecutStability \leftarrow 3;$
// Area threshold for MAE curve
$MAE\_areaThreshold \leftarrow estimateAdaptiveAreaThreshold(k\_values,$
                                   $MAE\_values,$
                                   $distanceBetweenTrianglePoints);$
// Slope threshold for MAE curve
$MAE\_slopeThreshold \leftarrow estimateAdaptiveSlopeThreshold(k\_values,$
                                   $MAE\_values);$
// Area threshold for curve with number of irrelevant clusters
$irrel\_areaThreshold \leftarrow 1.5;$
// Slope threshold for curve with number of irrelevant
   clusters (degrees)
$irrel\_slopeThreshold \leftarrow 22.5;$

// Variables to store best $k$ values for every curve
$best\_k\_values\_MAE[] \leftarrow ISAC(k\_values, MAE\_values,$
                          $distanceBetweenTrianglePoints,$
                          $consecutStability,$
                          $MAE\_areaThreshold, MAE\_slopeThreshold);$

$best\_k\_values\_irrel[] \leftarrow ISAC(k\_values, irrelCluster\_values,$
                          $distanceBetweenTrianglePoints,$
                          $consecutStability,$
                          $irrel\_areaThreshold, irrel\_slopeThreshold);$
$best\_k\_values[] \leftarrow commonValuesOn(best\_k\_values\_MAE,$
                                 $best\_k\_values\_irrel[]);$
```

- All daily consumption associated with non-household consumers has been removed. A non-domestic consumer is considered to be any consumer who, at any hour of any day, has a electricity consumption greater than 15 kWh.
- All observations where the consumption of any hour is missing have been removed.
- All observations with total daily consumption less than 100 W have been removed.

The final datasets to which it has been applied the methodology (see Section 3.1) are described below:

- Dataset from southeastern Spain was recorded from January 2020 to December 2021 for more than 3000 users. A total of 2396,741 observations were used after filtering data.
- Dataset from Ireland has been collected from approximately 4000 users over two years (2009–2010). A total of 2,522,976 observations were used after filtering data.

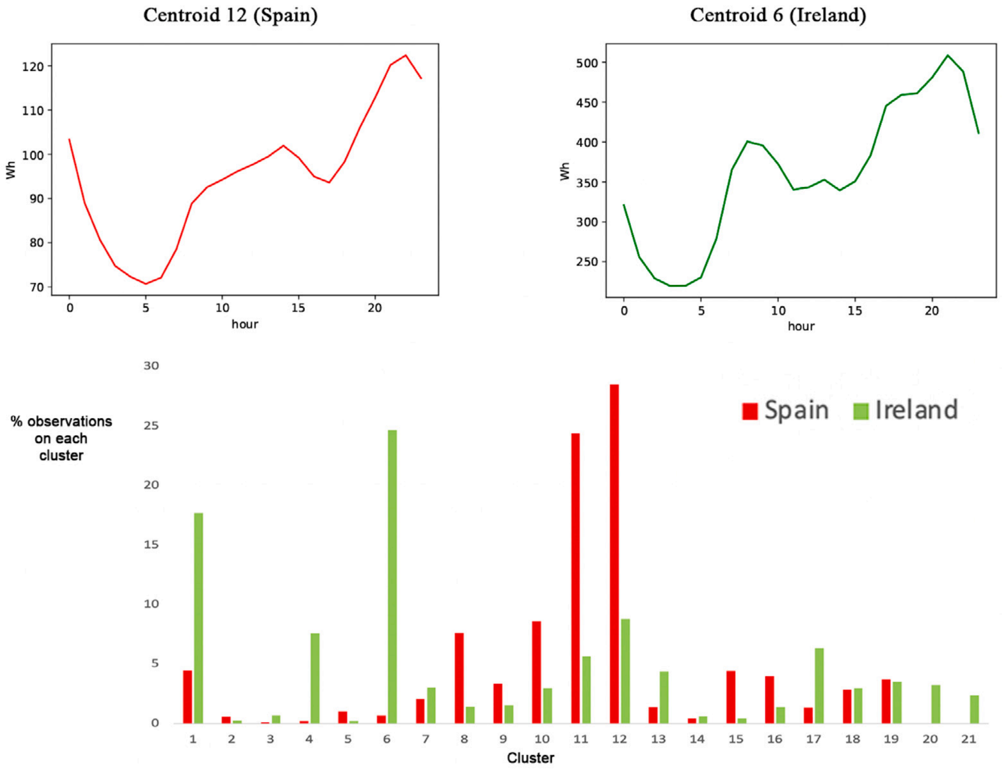


Fig. 6. The upper part of the figure shows the centroid representing the most frequently occurring cluster for the cities of Spain and Ireland. The lower part shows the percentage of observations for every cluster in each city.

4.2. Software and hardware used

The pre-processing data steps, the induction of clustering models, the calculation of metrics, the implementation of the ISAC method and k-ISAC_TLP procedure, and other computational tasks are executed mainly by using Python (Python Core Team, 2019), Spark (Zaharia et al., 2016), and MLlib (Machine Learning for Spark) (Meng et al., 2015).

The tool used for the implementation (scripts, proposed methodology, modeling, algorithms, calculation of metrics and visualization of results) was jupyter notebook.

The datasets used in the experimental design are big enough to limit the various clustering algorithms used. Only a few algorithms could carry out this task (k-means, BIRCH, DBSCAN, OPTICS, etc.), but finally, only two of them completed the task (k-means and bisecting k-means).

The experiments have been carried out using a MacBook Pro laptop with an Intel i7 processor (2.7 GHz) with 6 cores and 16 GB of RAM.

5. Results

The goal of the first phase is to carry out an exploratory search of data clusters. This search is guided by experts in the domain of electricity consumption. The experts defined a range of clusters between 10 and 33 ($k_{min} = 10$ and $k_{max} = 33$), where relevant models should be discovered. A smaller value could miss important groups, and a bigger value could unnecessarily delay the search (more than 30 or 35 groups and their profiles are difficult enough to be analyzed by experts).

In the first phase, several algorithms were applied to the datasets of both countries. Thus, several models were generated to identify consumption profiles using different values of number of clusters (k). Table 2 details the algorithms tested to induce clustering models. When the algorithm has some problem in the execution is indicated. This table also shows the processing times. The only algorithms capable of generating clustering models from both huge datasets have been k-means and bisecting k-means, both in parallel and non-parallel computing. The time reduction when using the parallel versions is remarkable.

Table 2
Summary of the clustering algorithms used in the experiments: execution time, use of parallelism, language and library.

	Algorithm	Time phase 1/2 (h)	Library/language
Non-parallelized algorithm	K-means	4.8/0.15	skicit-learn/Python
	Partitional (DTW)	Memory overflow	tsclust/R
	Partitional (GAK)	Memory overflow	tsclust/R
	Bisecting k-means	5.1/0.15	skicit-learn/Python
	BIRCH	Time exceeded	skicit-learn/Python
	DBSCAN	Time exceeded	skicit-learn/Python
	OPTICS	Time exceeded	skicit-learn/Python
Big data algorithm	K-means	0.5/0.15	MLib/Python (Spark)
	Bisecting k-means	0.6/0.15	MLib/Python (Spark)

In the second phase, the first stage is the selection of the algorithm. As described in Section 3.1, two metrics have been used for this task: the MAE and the Silhouette index. Lower MAE values are preferred, while higher Silhouette values are better.

For MAE curve, k-means and bisecting k-means algorithms obtain similar values, so the second criterion, the Silhouette index will help with the selection. This metric, calculated for both algorithms, shows that the clustering models obtained with k-means are better in terms of cohesion and separability of the clusters. Fig. 3 shows results for the Spanish dataset (the same behavior has been observed in Irish dataset).

The last stage in the second phase of the methodology is determining the number of clusters most appropriate for modeling the data with the proposed k-ISAC_TLP procedure. As explained in a previous section, the MAE and the number of small (and potentially useless) clusters present for different k values has been used.

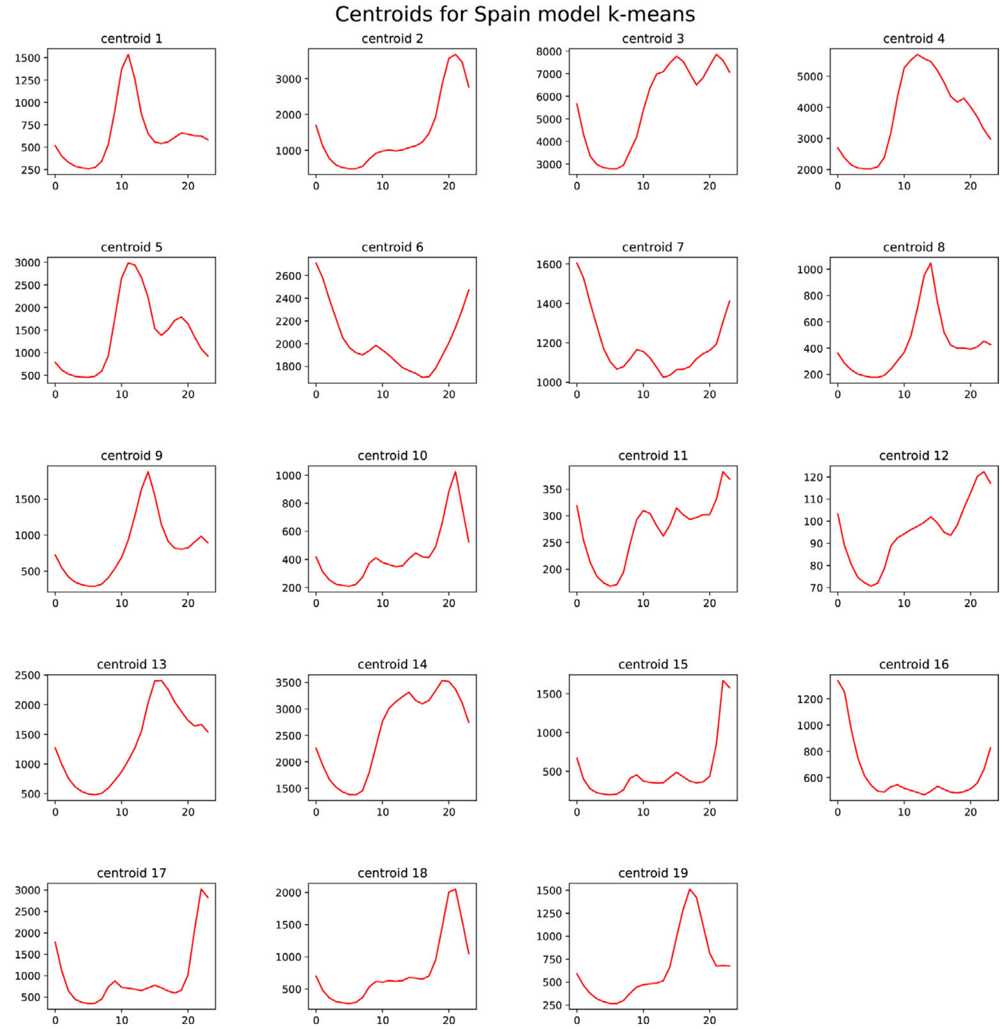


Fig. 7. Household electricity consumption profiles estimated from data of Spain. The X and Y-axis represent the time of day and the electricity consumption in Wh, respectively.

Executing the k-ISAC_TLP procedure on both datasets yields partly similar and partly different results. The difference relates to the shape of profiles detected, which are different for different types of users (which makes sense considering two datasets from different countries). The similarity lies in the number of profiles detected, that is similar. In both cases, the proposed k values are close to 20. Fig. 4 shows a visual description of the process to determine the candidate as an optimal number of clusters for Spain using the proposed k-ISAC_TLP procedure. The smallest common k value found is 19, so a model with 19 clusters is proposed for Spain. For the Irish dataset, shown in Fig. 5, the proposed k value is 21.

Once the most appropriate k values have been determined, the specific models induced are evaluated by the experts in the domain, to check the validity of the methodology and, in addition, to potentially discover novel knowledge.

Fig. 6 shows the percentage of observations in each cluster both the Spanish and Irish data and the load profile of the most common profiles in Spain and Ireland. The detailed electricity consumption profiles of the households identified with the proposed methodology are shown in Figs. 7–10. These consumption profiles are the centroids of the estimated clusters for each dataset. Figs. 7 and 8 describe Spanish profiles and Figs. 9 and 10 describe Irish profiles.

Although this research aims to present the methodology and not to describe each profile point-by-point, a summary of them is presented below. The objective is to prove the capabilities of the proposal, show some improvements achieved over other methodologies and highlight the advantages identified by the experts.

For Spanish data, more than 28% of the curves are in cluster 12; the consumption in this cluster ranges from 70 to 120 Wh with higher values after 20:00. Similar shape has the consumption profile of cluster 11, with more than 24% of curves; in this cluster, the consumption is higher and it ranges from 150 to 400 Wh. The consumption profiles of clusters 10, 15, 17, 18, and 19 have similar shapes although different values. They correspond to households where maximum consumption is around 20:00, this maximum value is 1000 Wh for cluster 10 (with 8.6% of curves), 1500, 3000, 2000, and 1500 Wh respectively for clusters 15, 17, 18, and 19. Clusters 1, 5, 8, 9, and 13 correspond to consumption profiles with a maximum around midday; the maximum consumption ranges between 1000 and 3000 Wh. Clusters 7 and 16 correspond to households where the maximum consumption is at 01:00 a.m. with a maximum consumption of 1600 and 1400 Wh respectively. The rest of the clusters have less than 1% of observations and therefore, they are not considered representative of any consumption profile.

For Irish data, more than 24% of observations are in cluster 6. The consumption in this cluster has a peak in the evening around 20:00 h with a consumption of 500 Wh and a consumption in the morning (but the peak is smaller) around 400 Wh. Cluster 9 have a similar shape but the values of consumption are greater, about 3000 Wh in the morning and 3300 Wh in the evening, only 1.5% of the observation belong to this cluster. The second group with more observations (almost 18%) is cluster 1. The shape of this profile has a consumption in the morning, about 8:00 h and then the consumption starts to increase at 15:00 with a maximum consumption around 20:00. Similar shapes have clusters

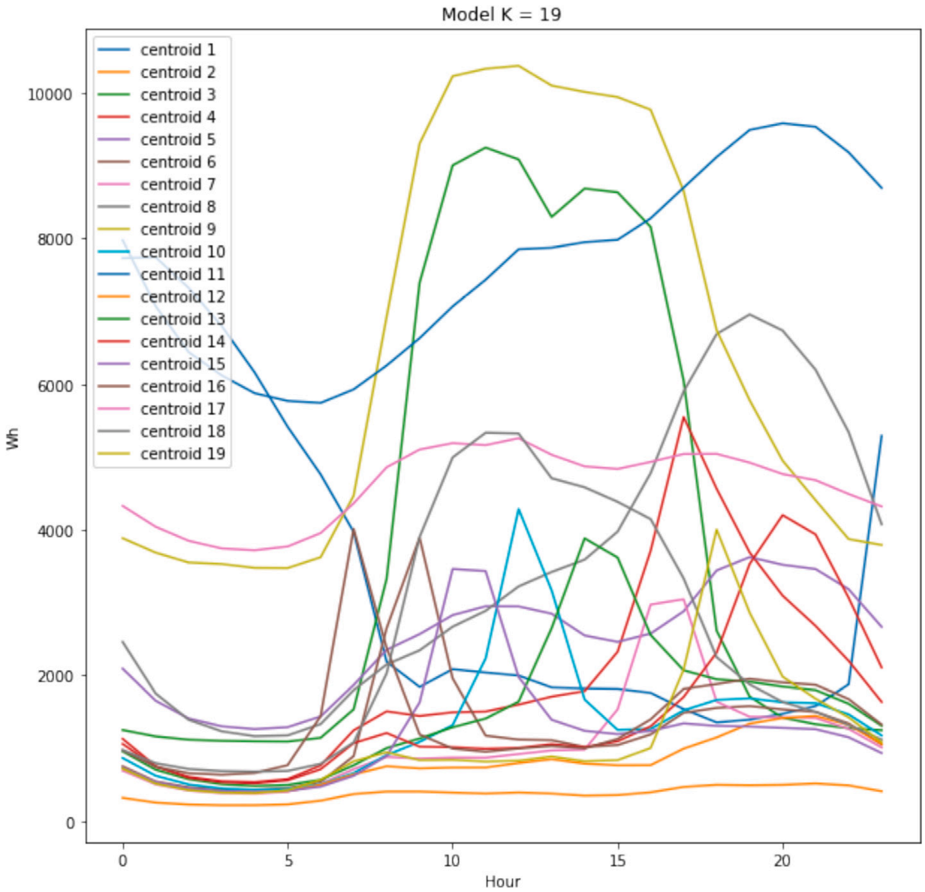


Fig. 8. Centroids for 19 types of consumptions profiles (Spain).

11, 17, 19, and 21, with a maximum consumption of 4000, 3000, 5000 and 6000 Wh respectively. The percentages of observations in these clusters vary between 2.4 and 6.4. Observations in cluster 4 have also consumption in the morning and a peak of consumption of 3000 Wh in the evening but at 17:00. Clusters 7, 10, 12, 13, 18, and 20 correspond to consumption profiles with a peak in the morning (between 2000 and 5000 Wh) and consumption in the evening (smaller than the peak, between 1000 and 2000 Wh); around the 25% of observations are in these clusters. Profile consumption of cluster 5 corresponds to consumption that happens in the night, with consumption of 8000 Wh; these observations correspond to non-domestic consumers. The rest of the clusters (3, 2, 8, 14, 15, and 16) have a profile of consumption along the day, from 9:00 (or 10:00) to 19:00 with values of consumption from 6000 to 9000 Wh; This consumption profile may be from businesses or companies and does not seem to correspond to domestic users.

As regards previous results obtained for data from Ireland, in McLoughlin et al. (2015) a part of the original dataset is used and a comparison can be made. They used data from six months for 3941 consumers. They propose a total of 10 different profile classes (PCs) of electricity load (or consumption) as a result of applying a series of clustering algorithms (k-means, SOM, k-medoids) using the Davies Bouldin index to determine the optimal number of clusters. Each one of these profile classes are defined by 48 values (half hourly electricity use).

Although their approach is different from the one proposed in our methodology, the dataset they use is smaller (in users and months), and the results are different (we find 21 profiles with 24 hourly values), the load profiles can be compared to assess similarities and differences. According to the PCs presented in Figure 6 of the previous work (McLoughlin et al., 2015), they seem to be coherent with those obtained in this work. Specifically, PCs 1, 4, and 9 that correspond to profiles with consumption less consumption in the morning and increases in the evening around 19:00 (or 20:00) are similar to clusters 1 and 6 obtained in our models. PCs 2 and 7 in McLoughlin et al. (2015) correspond to a profile consumption with the greatest consumption in the midday; similar profiles are in clusters 12, 13, and 18 in our analysis. In the same way, PCs 3 and 10 in that paper correspond to consumption with a peak in the morning and smaller consumption in the evening; similar profiles are in clusters 7, 10, and 20 in our analysis.

One important difference is that the clustering algorithms and the validation index have been applied in McLoughlin et al. (2015) to a series of daily consumption samples taken randomly from the original dataset. Besides the instability problem, sampling implies clustering models with a smaller number of clusters that cannot reflect the knowledge obtained by working with the entire dataset. Considering the whole dataset, as the methodology proposed in this paper does, can have a higher computational cost but gives a greater diversity of profiles. Using big data clustering algorithms allows the huge dataset to be processed entirely. Applying the proposed methodology allows for generating models that detect more electricity consumption profiles than other methodologies. For example, profiles associated with non-domestic consumers (businesses or companies) are detected.

In addition to not having to sample, some additional advantages observed by experts are related to the low level of pre-processing carried out:

- Only some filters are defined to eliminate incomplete or extremely odd observations.
- The normalization step, usually included in many methodologies, is not applied here. This can reveal that some profiles, whose shape can be similar but with different consumption ranges, are related to the same behavior of electricity use. The differences

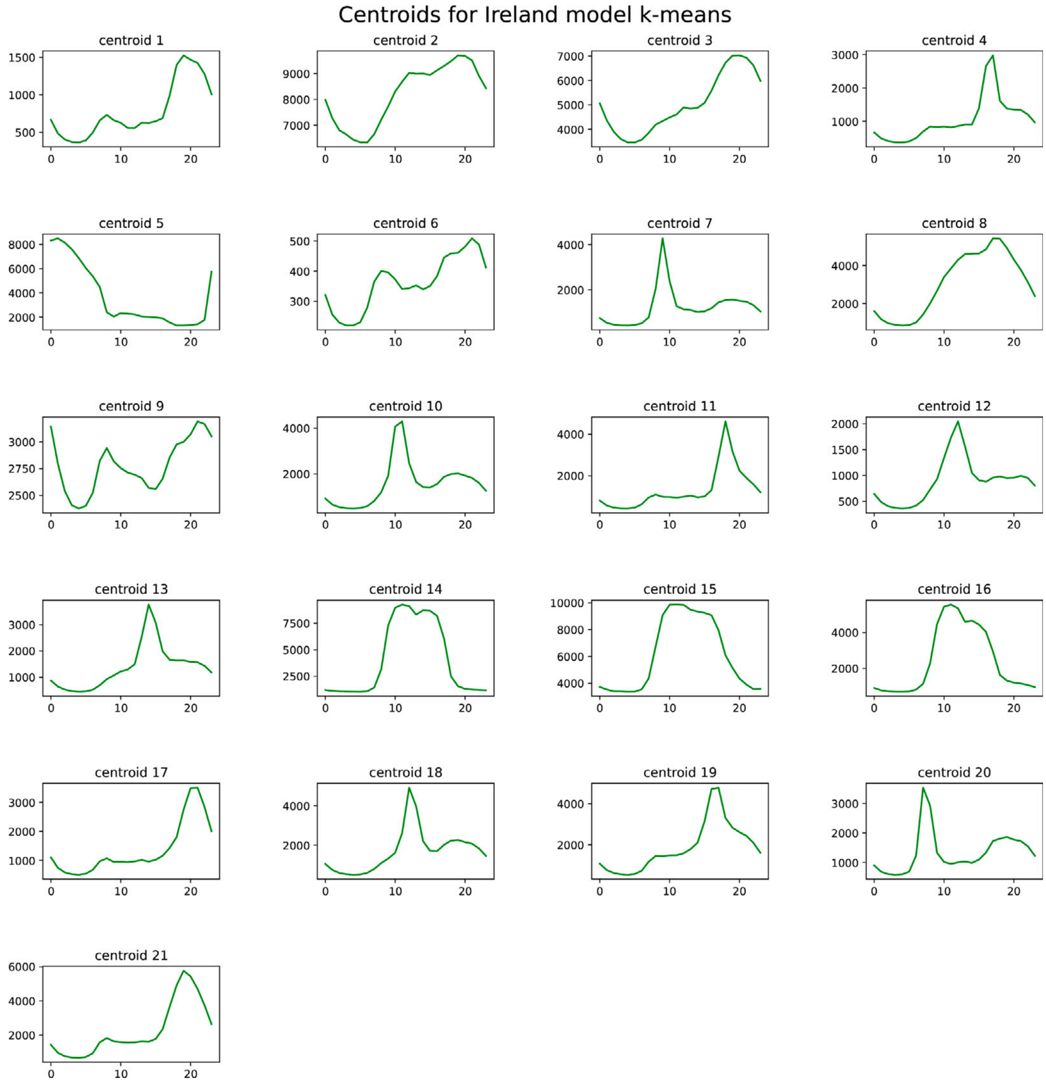


Fig. 9. Household electricity consumption profiles estimated from data of Ireland. The X and Y-axis represent the time of day and the electricity consumption in Wh, respectively.

derived from the range can be related to the household configuration (number of members or type of appliances) and can be a source of study and improvement. An additional advantage is the absence of the denormalization step, a complex task that can insert some errors into the process.

- The segmentation process of the dataset, usually performed to apply a cascading clustering, is not applied here. This idea reduces the bias introduced by the segmentation and does not force the repetition of several clustering processes in isolated subsets. This approach also makes comprehension of the profiles easier because there is no need to analyze all the clusters for all the segmented subsets.

Regarding the other works using the Ireland data, in Rajabi et al. (2020), both the clustering of one residential customer and the clustering of a large number of users are analyzed. For this second type, they found 16 different types of clusters for weekdays and another 16 for weekends. It is not possible to compare the results directly because they use half-hour recordings. In Sun et al. (2020), although they also use the data from Ireland, data are sampled only over the period 1st December to 31st December. For this period, they found 6 typical daily load profiles on weekends and 6 on weekdays for half-hour data. Finally, in Guo et al. (2022), both household characteristics and historical and contemporary electricity usage data are used to segment consumers.

6. Conclusions and future works

This paper has presented a new methodology for discovering household electricity consumption profiles. In the first phase, it is essential to delimit a range for the search space in which the number of clusters must be found. Subsequently, multiple clustering algorithms can be used. Still, a crucial point is the definition of the measures to be optimized, considering the domain (such as the accumulated consumption error in the model or the number of irrelevant clusters). Its main novel aspects arise from close work with experts in the field. An important point is the incorporation of the error metric as a criterion to complement the clustering validity indices. Another highlight of this methodology is a new procedure to automatically search for the most appropriate number of clusters in the range defined by the experts.

This simple methodology has been implemented as open-source software and has proven helpful in helping experts learn from huge datasets by minimizing pre-processing steps. Considering the entire dataset, induced models avoid the usual sampling instability. By working with the complete data set, profiles are revealed that would not be discovered using sampling. By removing the normalization step, the complex denormalization task is not needed. At the same time, avoiding normalization, all the profiles can be compared directly, and profiles with common shapes but different ranges are identified. By dispensing with data pre-processing, which segments the dataset into

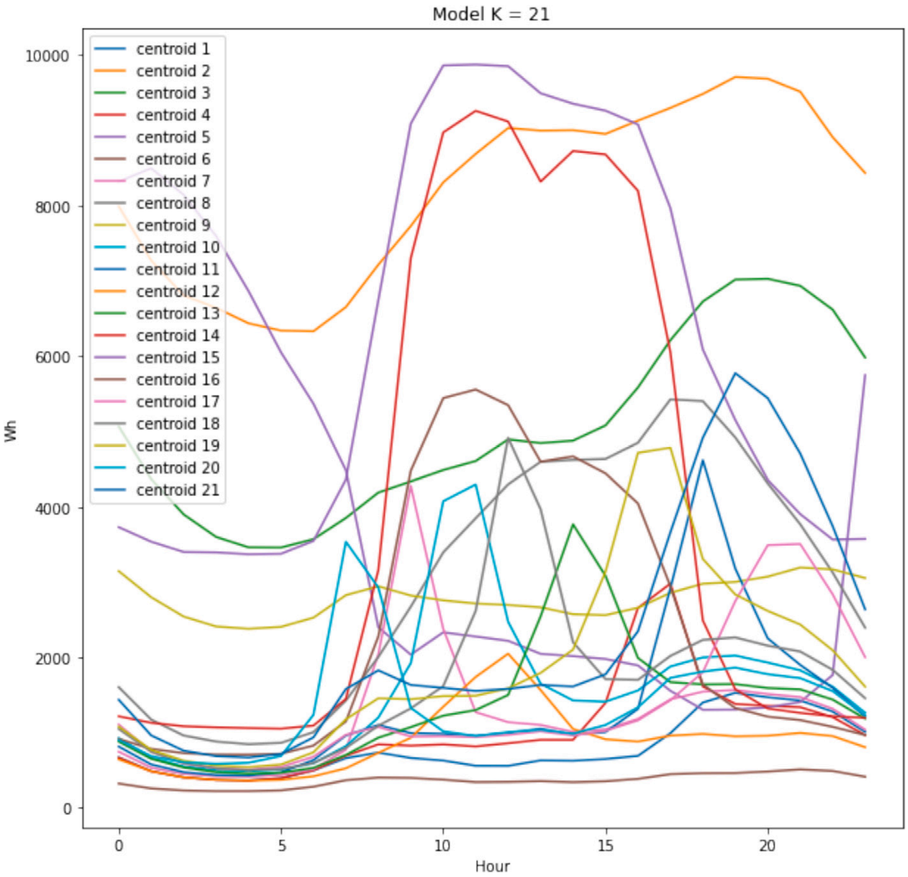


Fig. 10. Centroids for 21 types of consumptions profiles (Ireland).

different types of days, months or seasons, this methodology identifies the profiles regardless of these temporal aspects.

The models obtained help experts quickly gain new knowledge in the field of household electricity consumption. They also can directly impact the fight against climate change and the improvement of urban sustainability.

The methodology is easily extensible to problems of any domain where clustering algorithms are applicable and experts can put their knowledge to good use. The proposed procedure to identify an appropriate number of clusters is also extensible to other fields and will be the subject of future work.

Further research will make it possible to assign defining characteristics to each profile. Those characteristics may not be unique because the same profile can be repeated in different seasons, months or days for the same or different users. Therefore, the typical load profile classification for a user on a specific day will be more flexible.

One of the main limitations of the proposed methodology is that it can become inefficient for very wide ranges of k, since it is necessary to generate a model for each value of k with each of the proposed algorithms and calculate the metrics for each of them. A more intelligent search for the optimal number of clusters, without having to calculate models for all possible values, would be useful.

Another limitation is the difficulty of finding suitable parameters of the ISAC algorithm depending on the application domain. Therefore, it would also be desirable to work on the proposal and analysis of new curves, metrics and parameter configurations of the ISAC algorithm in order to obtain a general configuration valid for a wide variety of domains.

Reproducibility and released software

To contribute by sharing the proposed methodology, we provide the code that implements the methodology. In https://github.com/ursusdm, in the repository called consumptionprofiles, we distribute: (1) scripts with the proposed methodology, (2) scripts with a new procedure to determine appropriate numbers of clusters (k), and (3) instructions with the steps to follow.

For privacy reasons, sharing the datasets from Spain has not been possible, but the Irish dataset is available (Commission for Energy Regulation (CER), 2012).

CRediT authorship contribution statement

Francisco Rodríguez-Gómez: Methodology, Software, Validation, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization. **José del Campo-Ávila:** Conceptualization, Methodology, Validation, Investigation, Writing – original draft, Writing – review & editing, Visualization, Supervision. **Llanos Mora-López:** Conceptualization, Methodology, Validation, Resources, Writing – original draft, Writing – review & editing, Supervision, Project administration.

Declaration of competing 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.

Data availability

One dataset is available if asked to owner. Another dataset is not public. More details are given in the paper.

Acknowledgments

The authors would like to thank the University College Dublin Library the access to the Irish Social Science Data Archive (ISSDA). This work was supported by Grant RTI2018-095097-B-I00 funded by MCIN (Spain), Grant CPP2021-008403 funded by MCIN/AEI/10.13039/501100011033 and by the "European Union NextGenerationEU/PRTR". Funding for open access charge: Universidad de Málaga/CBUA.

Appendix A. Auxiliary algorithms

Algorithm 3: estimateAdaptiveAreaThreshold

```
Input: // DATA: values for $k$ to be considered in the range
        // defined by the experts
        $k\_values[k_{min}, k_{max}] = k_{min}, k_{min} + 1, \ldots, k_{max}$
        // DATA: List with MAE values measured for different $k$
          values. For simplicity, assume that indexes are
          named $k_{min}, k_{min} + 1, \ldots, k_{max}$
        $curve\_values[k_{min}, k_{max}]$
        // number of values in X-axis between points of each
          triangle
        $distanceBetweenTrianglePoints$
Output: // Area threshold
         $areaThreshold$

// Largest triangle in the curve
$p'_x \leftarrow k_{min}$ ; $p'_y \leftarrow curve\_values[p'_x]$;
$q'_x \leftarrow \lceil (k_{min} + k_{max})/2 \rceil$ ; $q'_y \leftarrow curve\_values[q'_x]$;
$r'_x \leftarrow k_{max}$ ; $r'_y \leftarrow curve\_values[r'_x]$;
// Triangle proportional to largest triangle
// Share middle vertex ($q'_x$). Change extreme vertices ($p''_x$ and $r''_x$)
$p''_x \leftarrow q'_x - distanceBetweenTrianglePoints$;
$p''_y \leftarrow q'_y + ((p'_y - q'_y) \cdot distanceBetweenTrianglePoints)/(q'_x - p'_x)$;
$r''_x \leftarrow q'_x + distanceBetweenTrianglePoints$;
$r''_y \leftarrow q'_y + ((r'_y - q'_y) \cdot distanceBetweenTrianglePoints)/(r'_x - q'_x)$;
$areaThreshold \leftarrow calculateArea(p'', q', r'')$;
```

Algorithm 4: estimateAdaptiveSlopeThreshold

```
Input: // DATA: values for $k$ to be considered in the range
        // defined by the experts
        $k\_values[k_{min}, k_{max}] = k_{min}, k_{min} + 1, \ldots, k_{max}$
        // DATA: List with curve values measured for different $k$
          values
        // For simplicity, assume that indexes are named
          $k_{min}, k_{min} + 1, \ldots, k_{max}$
        $curve\_values[k_{min}, k_{max}]$
Output: // Slope threshold
         $slopeThreshold$

// Largest triangle in the curve
$p'_x \leftarrow k_{min}$ ; $p'_y \leftarrow curve\_values[p'_x]$;
$r'_x \leftarrow k_{max}$ ; $r'_y \leftarrow curve\_values[r'_x]$;
$slopeThreshold \leftarrow calculateSlope(p', r')$;
```

References

Aggarwal, C.C., 2015. Data Mining. Springer International Publishing, Cham, http://dx.doi.org/10.1007/978-3-319-14142-8.

Ankerst, M., Breunig, M.M., Kriegel, H.-P., Sander, J., 1999. OPTICS: ordering points to identify the clustering structure. In: Proceedings on 1999 ACM SIGMOD International Conference on Management of Data, Vol. 28. ACM PUB27, New York, NY, USA, pp. 49–60. http://dx.doi.org/10.1145/304181.304187.

Berahmand, K., Bouyer, A., Vasighi, M., 2018. Community detection in complex networks by detecting and expanding core nodes through extended local similarity of nodes. IEEE Trans. Comput. Soc. Syst. 5 (4), 1021–1033. http://dx.doi.org/10.1109/TCSS.2018.2879494.

Borjigin, S., Guo, C., 2013. Non-unique cluster numbers determination methods based on stability in spectral clustering. Knowl. Inf. Syst. 36 (2), 439–458. http://dx.doi.org/10.1007/S10115-012-0547-0/TABLES/14.

Castillo, G., Gama, J., 2006. An adaptive prequential learning framework for Bayesian network classifiers. Lecture Notes in Artificial Intelligence 4213, 67–78. http://dx.doi.org/10.1007/11871637_11.

Cembranel, S.S., Lezama, F., Soares, J., Ramos, S., Gomes, A., Vale, Z., 2019. A short review on data mining techniques for electricity customers characterization. In: 2019 IEEE PES GTD Grand International Conference and Exposition Asia (GTD Asia). IEEE, pp. 194–199. http://dx.doi.org/10.1109/GTDAsia.2019.8715891.

Charrad, M., Ghazzali, N., Boiteau, V., Niknafs, A., 2014. Nbclust: An R package for determining the relevant number of clusters in a data set. J. Stat. Softw. 61 (6), 1–36. http://dx.doi.org/10.18637/jss.v061.i06.

Commission for Energy Regulation (CER), 2012. CER Smart Metering Project - Electricity Customer Behaviour Trial, 2009–2010, first ed. Irish Social Science Data Archive, URL: https://www.ucd.ie/issda/data/commissionforenergyregulationcer/.

Cuturi, M., 2011. Fast global alignment kernels. In: Proceedings of the 28th International Conference on Machine Learning, ICML 2011. pp. 929–936.

Dafir, Z., Lamari, Y., Slaoui, S.C., 2021. A survey on parallel clustering algorithms for Big Data. Artif. Intell. Rev. 54 (4), 2411–2443. http://dx.doi.org/10.1007/s10462-020-09918-2.

Davies, D.L., Bouldin, D.W., 1979. A cluster separation measure. IEEE Trans. Pattern Anal. Mach. Intell. PAMI-1 (2), 224–227. http://dx.doi.org/10.1109/TPAMI.1979.4766909.

del Campo-Ávila, J., Ramos-Jiménez, G., Gama, J., Morales-Bueno, R., 2008. Improving the performance of an incremental algorithm driven by error margins. Intell. Data Anal. 12 (3), 305–318. http://dx.doi.org/10.3233/ida-2008-12305.

Dhanapal, J., Perumal, T., 2016. Inflated power iteration clustering algorithm to optimize convergence using Lagrangian constraint. In: Advances in Intelligent Systems and Computing, Vol. 465. Springer Verlag, pp. 227–237. http://dx.doi.org/10.1007/978-3-319-33622-0_21.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In: KDD'96: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. pp. 226–231.

Ezugwu, A.E., Shukla, A.K., Agbaje, M.B., Oyelade, O.N., José-García, A., Agushaka, J.O., 2021. Automatic clustering algorithms: a systematic review and bibliometric analysis of relevant literature. Neural Comput. Appl. 33 (11), 6247–6306. http://dx.doi.org/10.1007/s00521-020-05395-4.

Figueiredo, V., Duarte, F.J., Rodrigues, F., Vale, Z., Gouveia, J., 2003. Electric energy customer characterization by clustering. In: IEEE Intelligent Systems Applications to Power Systems. p. 6.

Guo, Z., O'Hanley, J.R., Gibson, S., 2022. Predicting residential electricity consumption patterns based on smart meter and household data: A case study from the Republic of Ireland. Util. Policy 79, 101446. http://dx.doi.org/10.1016/j.jup.2022.101446.

Hamerly, G., Elkan, C., 2003. Learning the k in k-means. In: Advances in Neural Information Processing Systems (NIPS). MIT Press, pp. 281–288.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning: Data Mining, Inference and Prediction, second ed. Springer.

Kaur, R., Gabrijelčič, D., 2022. Behavior segmentation of electricity consumption patterns: A cluster analytical approach. Knowl.-Based Syst. 251, 109236. http://dx.doi.org/10.1016/j.knosys.2022.109236.

Kwac, J., Flora, J., Rajagopal, R., 2014. Household energy consumption segmentation using hourly data. IEEE Trans. Smart Grid 5 (1), 420–430. http://dx.doi.org/10.1109/TSG.2013.2278477.

Lin, F., Cohen, W.W., 2010. Power iteration clustering. In: Proceedings of the 27th International Conference on International Conference on Machine Learning. ICML '10, Omni Press, Madison, WI, USA, pp. 655–662.

MacQueen, J., 1967. Some methods for classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. pp. 281–297.

Martinez-Plumed, F., Contreras-Ochando, L., Ferri, C., Hernandez-Orallo, J., Kull, M., Lachiche, N., Ramirez-Quintana, M.J., Flach, P., 2021. CRISP-DM twenty years later: From data mining processes to data science trajectories. IEEE Trans. Knowl. Data Eng. 33 (8), 3048–3061. http://dx.doi.org/10.1109/TKDE.2019.2962680.

McLoughlin, F., Duffy, A., Conlon, M., 2015. A clustering approach to domestic electricity load profile characterisation using smart metering data. Appl. Energy 141, 190–199. http://dx.doi.org/10.1016/j.apenergy.2014.12.039.

Meng, X., Bradley, J., Yavuz, B., Sparks, E., Venkataraman, S., Liu, D., Freeman, J., Tsai, D., Amde, M., Owen, S., Xin, D., Xin, R., Franklin, M.J., Zadeh, R., Zaharia, M., Talwalkar, A., 2015. MLlib: Machine learning in apache spark. http://dx.doi.org/10.48550/ARXIV.1505.06807.

Mets, K., Depuydt, F., Develder, C., 2016. Two-stage load pattern clustering using fast wavelet transformation. IEEE Trans. Smart Grid 7 (5), 2250–2259. http://dx.doi.org/10.1109/TSG.2015.2446935.

Pelleg, D., Moore, A.W., 2000. X-means: Extending K-means with efficient estimation of the number of clusters. In: Proceedings of the Seventeenth International Conference on Machine Learning. pp. 727–734.

Python Core Team, 2019. Python: A Dynamic, Open Source Programming Language. Python Software Foundation, URL: https://www.python.org/.

Rafiq, H., Manandhar, P., Rodriguez-Ubinas, E., Barbosa, J.D., Qureshi, O.A., 2023. Analysis of residential electricity consumption patterns utilizing smart-meter data: Dubai as a case study. Energy Build. 291, 113103. http://dx.doi.org/10.1016/j.enbuild.2023.113103.

Raj, S.A.P., Vidyaathulasiraman, 2021. Determining optimal number of K for e-learning groups clustered using K-medoid. Int. J. Adv. Comput. Sci. Appl. 12 (6), 400–407. http://dx.doi.org/10.14569/IJACSA.2021.0120644.

Rajabi, A., Eskandari, M., Ghadi, M., Li, L., Zhang, J., Siano, P., 2020. A comparative study of clustering techniques for electrical load pattern segmentation. Renew. Sustain. Energy Rev. 120, http://dx.doi.org/10.1016/j.rser.2019.109628.

Räsänen, T., Voukantsis, D., Niska, H., Karatzas, K., Kolehmainen, M., 2010. Data-based method for creating electricity use load profiles using large amount of customer-specific hourly measured electricity use data. Appl. Energy 87 (11), 3538–3545. http://dx.doi.org/10.1016/j.apenergy.2010.05.015.

Rostami, M., Oussalah, M., Berahmand, K., Farrahi, V., 2023. Community detection algorithms in healthcare applications: A systematic review. IEEE Access 11, 30247–30272. http://dx.doi.org/10.1109/ACCESS.2023.3260652.

Rousseeuw, P.J., 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. 20, 53–65. http://dx.doi.org/10.1016/0377-0427(87)90125-7.

Saeed, M.M., Aghbari, Z.A., Alsharidah, M., 2020. Big data clustering techniques based on Spark: a literature review. PeerJ Comput. Sci. 6, 1–28. http://dx.doi.org/10.7717/PEERJ-CS.321.

Sakoe, H., Chiba, S., 1978. Dynamic programming algorithm optimization for spoken word recognition. IEEE Trans. Acoust. Speech Signal Process. 26 (1), 43.

Sharan, R., Maron-Katz, A., Shamir, R., 2003. CLICK and EXPANDER: A system for clustering and visualizing gene expression data. Bioinformatics 19 (14), 1787–1799. http://dx.doi.org/10.1093/bioinformatics/btg232.

Shi, C., Wei, B., Wei, S., Wang, W., Liu, H., Liu, J., 2021. A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm. EURASIP J. Wireless Commun. Networking 2021 (1), 31. http://dx.doi.org/10.1186/s13638-021-01910-w.

Steinbach, M., Karypis, G., Kumar, V., 2000. A comparison of document clustering techniques. In: KDD Workshop on Text Mining. pp. 1–20.

Sun, L., Zhou, K., Yang, S., 2020. An ensemble clustering based framework for household load profiling and driven factors identification. Sustainable Cities Soc. 53, http://dx.doi.org/10.1016/j.scs.2019.101958.

Toussaint, W., 2019. Domestic electrical load metering, hourly data 1994–2014 [dataset]. Version 1. http://dx.doi.org/10.25828/56nh-fw77.

Toussaint, W., Moodley, D., 2020. Clustering residential electricity consumption data to create archetypes that capture household behaviour in South Africa. S. Afr. Comput. J. 32, 1–34. http://dx.doi.org/10.18489/sacj.v32i2.845.

Verdu, S., Garcia, M., Franco, F., Encinas, N., Marin, A., Molina, A., Lazaro, E., 2004. Characterization and identification of electrical customers through the use of self-organizing maps and daily load parameters. In: IEEE PES Power Systems Conference and Exposition, 2004. Vol. 2. IEEE, pp. 1240–1247. http://dx.doi.org/10.1109/PSCE.2004.1397641.

Xu, D., Tian, Y., 2015. A comprehensive survey of clustering algorithms. Ann. Data Sci. 2 (2), 165–193. http://dx.doi.org/10.1007/s40745-015-0040-1.

Yilmaz, S., Chambers, J., Patel, M., 2019. Comparison of clustering approaches for domestic electricity load profile characterisation - Implications for demand side management. Energy 180, 665–677. http://dx.doi.org/10.1016/j.energy.2019.05.124.

Zaharia, M., Xin, R.S., Wendell, P., Das, T., Armbrust, M., Dave, A., Meng, X., Rosen, J., Venkataraman, S., Franklin, M.J., Ghodsi, A., Gonzalez, J., Shenker, S., Stoica, I., 2016. Apache Spark: A unified engine for big data processing. Commun. ACM 59 (11), 56–65. http://dx.doi.org/10.1145/2934664.

Zhang, T., Ramakrishnan, R., Livny, M., 1996. BIRCH: an efficient data clustering method for very large databases. ACM SIGMOD Rec. 25 (2), 103–114. http://dx.doi.org/10.1145/235968.233324.